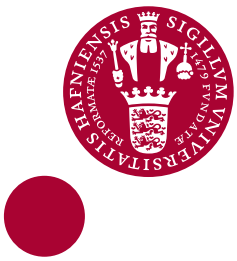


The Design and Implementation of an Assessment Method Combining Formative and Summative Use of Assessment

Nielsen, Sanne Schnell; Dolin, Jens; Bruun, Jesper; Jensen, Sofie Birch

Published in:
Evaluation and assessment of student learning and development: Part 11, Strand 11

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

PART 11: STRAND 11

Evaluation and Assessment of Student Learning and Development

Co-editors: *Jens Dolin*

We acknowledge the considerable contributions of our colleague Per Morten Kind (RIP) as Co-Chair of Strand 11.

CONTENTS

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

STRAND 11: INTRODUCTION

EVALUATION AND ASSESSMENT OF STUDENT LEARNING AND DEVELOPMENT

Strand 11 focused on a wide range of issues related to the evaluation and assessment of student learning and development, both on an individual and classroom level and on a national and international level.

The *use* of assessment instruments (like tests, questionnaires etc.), as tools for exploring and answering other research questions of interest, is *not* part of the strand mission. In Strand 11 the focus is primarily on the instrument itself, the emphasis is on the development, implementation, validation and use of assessment instruments – and on the consequences of the use of the instruments. These can include standardized tests, achievement tests, high stakes tests, and instruments for measuring attitudes, interests, beliefs, self-efficacy, science process skills and competences, conceptual understandings, and so on. They may be developed with a view to making assessment more 'authentic' in some sense, to facilitate formative use of assessment, or to improve summative assessment of student learning.

Jens Dolin

(greatly mourning the loss of Per Morten Kind who participated in the strand work even heavily marked by his illness)

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

USE OF CASE STUDY TO DEVELOP AND EXEMPLIFY OF A MODEL OF TEACHER ASSESSMENT

Sarah Earle
Bath Spa University, England

The Teacher Assessment in Primary Science (TAPS) project is based at Bath Spa University and funded by the Primary Science Teaching Trust. Using a Design-Based Research (DBR) approach it has worked collaboratively with schools to operationalize a model of teacher assessment put forward by the Nuffield Foundation (2012) whereby formative classroom assessment information is summarized for summative purposes (Davies, Earle, McMahon, Howe & Collier, 2017). This paper presents a case study of one of the TAPS project schools utilizing data from school visits and TAPS development days, collected between June 2013 and June 2015. The case study addresses research questions around the nature of formative and summative assessment, and the relationship between the two within the school. In discussion of the case, the aim is to explore the enactment of a 'formative to summative' approach to assessment in primary science, as proposed by the Nuffield Foundation (2012) and TAPS (Davies et al. 2016). Key features of practice drawn from the case include the use of: pupil self and peer assessment; explicit criteria; and whole school moderation meetings. Questions are raised for the DBR approach regarding benefits of the partnership for the school, since little change in primary science practice was seen during the case study period; whilst the school has supported development of the TAPS pyramid model, for example, with the inclusion of a 'shared understanding' criterion in response to practice seen in schools.

Keywords: formative assessment, design-based research, primary school

INTRODUCTION

Assessment is fundamental to the practice of education, yet it is not neutral, it is value-laden; assessment processes determine what is valuable to learn and what success will look like, it: *"creates and shapes what is measured"* (Stobart, 2008: 1). Since assessment shapes the curriculum as experienced by children; it is essential for such assessment practices to be well understood by teachers. The functions and effect of assessment have received much attention, with some arguing (Black & Wiliam, 1998) that assessment should have an impact on learning otherwise there is little point in conducting the assessment in the first place. Research into formative assessment champions the use of assessment to support learners with their next steps (Gardner, Harlen, Hayward, Stobart & Montgomery, 2010); whilst summative assessment became viewed in a negative light because of suggestions that it was the cause of curriculum narrowing and teaching to the test (Harlen, 2013). However, education systems require both purposes to be fulfilled, with assessment information used to support learning and to summarise achievements. Such a clash between a positive view of formative assessment and a negative view of summative assessment may be counter-productive, leading teachers to run separate, and consequently unmanageable, assessment systems (Earle, 2014).

A closer relationship between formative and summative assessment is seen by some as crucial to effective teacher assessment (Harlen, 2013; Hodgson & Pyle, 2010; Nuffield Foundation, 2012; Wiliam & Black, 1996). An expert group convened by the Nuffield Foundation (2012) proposed that assessment information gathered for the purposes of formative assessment during

the course of typical classroom activities could also be used to serve summative purposes by informing summaries of pupil performance when reporting for different purposes. Their pyramid-shaped model of teacher assessment, in which information flows from the classroom base to the reporting tip, was developed in response to growing concerns for the negative impact of external summative testing, skewing the taught curriculum to that which was easily tested (Gardner et al., 2010). Whilst the Nuffield model was welcomed by the primary science community, it contained little detail of how to implement its proposals.

The Teacher Assessment in Primary Science (TAPS) project, based at Bath Spa University and funded by the Primary Science Teaching Trust (PSTT), set out to operationalize the Nuffield proposals using a Design-Based Research (DBR) approach to work collaboratively with project schools to translate research into practice (Anderson & Shattuck, 2012). During iterative cycles the TAPS school self-evaluation model was developed, containing criteria and examples to support schools to develop their assessment processes (Earle, McMahon, Howe, Collier & Davies, 2016). This paper presents a case study of one of the TAPS project schools, with the aims of testing and exemplifying the model of 'formative to summative' assessment.

METHODS

The case was selected as a 'critical' or 'instrumental' case (Stake, 2006), to provide a test case for the 'formative to summative' model, since School A was an award-winning PSTT school who asserted that they used formative classroom assessments to make summative judgements,. In this paper, the following research questions are addressed:

RQ1. What are the characteristics of **formative** teacher assessment in science in School A?

RQ2. What are the characteristics of **summative** teacher assessment in science in School A?

RQ3. What is the **relationship between** formative and summative assessment in School A?

The data for School A was collected, with ethical agreement from school and participants, between June 2013 and June 2015 and consisted of: non-participant lesson observations (N=3) and interviews (N=3) from six school visits; school documentation including policies, lesson plans and assessment records; and activities from six TAPS development days. On many occasions the school was represented by the science subject leader, the class teacher responsible for leading the development of science across the school, thus much of the data was from her perspective. In order to triangulate the reported practice, classroom observations and documentation from across the school were included in the data. Particular attention was paid to the research questions in order that the analysis should remain focused on the relationship between formative and summative assessment, whilst placing this within the rich context of the school. Qualitative data analysis was supported by ATLAS.ti software, with codes developed deductively from the research questions and the TAPS pyramid (Earle et al., 2016), and inductively from the data itself. 'Higher order codes', those which were both frequently represented in the data-set and pertinent to the research questions (Bryman, 2016), have been selected as examples for discussion in this brief paper.

RESULTS

Formative teacher assessment

Formative assessment at School A was built into lesson planning, for example in the form of key questions for teachers to ask their pupils. Classroom discussion was a prominent feature in all three of the observed lessons (see Table 1), with each teacher used strategies like talk partners to increase participation and wait time. For example, in the Year 6 lesson, pair talk dominated with the teacher 'listening in' to discussions to support her formative assessment, then asking probing questions to stimulate further discussion. In the Year 5 lesson, it was noted that the teacher was 'withholding judgement' (Table 1, row 4) during the class discussion. The teacher questioning focused on explanations and use of vocabulary, but the teacher did not say 'that's right' and move on. This could support a more dialogic (Alexander, 2008) approach to discussion, moving beyond the mere 'call and response' of interactive-authoritative dialogue (Mortimer & Scott, 2003). By withholding summative judgement of pupils' answers, the children were prompted to explain further and the teacher received richer formative assessment information, from a greater number of children.

Table 1. Lesson observation field notes organised using TAPS pyramid Teacher layer criteria.

Lesson	**Y4 Keys lesson** Creating post-it keys to categorise animals	**Y5 Earth in space lesson** Using balls to model orbit of Earth	**Y6 Inheritance lesson** Exploring inherited characteristics in dogs and own families
Date	March 2014	January 2014	January 2014
Teachers involve students in discussing **learning goals** and standards	*Raised hands to show if find keys tricky.* *Mini-plenary to look at others' work – what do you notice?*	*Importance of using science vocab* *4 or 5 'star'scientists*	(did not observe start of lesson)
Teachers gather evidence of their students' learning through **questioning/ discussion**	*Discussed kind of Qs in branching database.* *Open Qs for talk partners: What hab in sch? What is it like? – asked for more detail*	*Qs emphasising expl - probed explanations and meaning/use of vocab - Withhold judgement so ch have to expl for selves*	*Probed children's meaning of inheritance vocab* *'No hands up' strategy*
Teachers gather evidence of their students' learning through **observation**	*Groups building post-it keys – spotted clearest and pointed children in that direction*	*Observe groups modelling Earth orbiting*	*Pairs recording ideas on whiteboards while teacher circulates*
Teachers gather evidence of their students' learning	*Post-it branching database* *Assessment notes on plans for children that*	*Look at group's explanation and modelling.*	*Whiteboards to note family characteristics. Written explanation of*

through **study of products**	*stand out – above or below*	*Draw and write explanation.*	*what learnt and examples*
Teachers use assessment to advance students' learning by **adapting** the pace, challenge and content of activities	*Previous lesson found branching keys difficult so doing in mixed ability groups* *Pupils identify how to improve their key e.g. missing Y/N*	*Physically modelling Earth's orbit in a circle since virtual experiment looks like oval. Did not move onto day and night since challenged enough by orbit whilst spinning.*	*Provides word to help explain e.g. characteristics, structure for recording: Mum, Dad, me.*
Teachers use assessment to advance students' learning by **giving feedback**	*Go around groups to check on clarity of Qs*	*Asking if can use better science words*	*Say more than 'face' for characteristics*
Teachers use assessment to advance students' learning by providing time for students to **reflect on and assess their own work**	*Evaluating Qs* *Pairs walk around and look at other's work – what notice?* *Return to own keys and improve*	*4th child in group to listen and watch – are they explaining using sci vocab, watch groups and give feedback, decide if 4 or 5 'star' scientists and write in margin*	(did not observe end of lesson)

Pupil self and peer assessment was supported by explicit success criteria, with stages of scientific inquiry displayed on the wall and referred to in lessons. For example, the subject leader's Year 5 lesson (January 2014) began with a whole class carpet discussion about the Earth and sun. In the main part of the lesson the pupils worked in pairs or threes to physically model the orbit of the Earth around the sun using different sized balls. As the children moved the 'Earth ball' they gave a commentary on what was happening, which was then peer-assessed for clarity and accuracy, with the groups giving advice to each other for how to improve their explanations. The teacher emphasized the accurate use of scientific vocabulary, pointing to the success criteria on the wall, leading the pupils to listen out for the word 'orbit' or 'axis' in the explanations. The use of explicit success criteria, a key feature of formative assessment (Wiliam, 2011), supported both teacher and pupil assessment in the observed lessons.

Pupil recording included 'floor books' for younger children in the school (a large-format, 'home-made' book), where an adult scribed their responses verbatim, whilst older children made focused recordings in their science books. The subject leader stated that:

> "*Marking is used to feed judgements back to children. Children are given the opportunity to respond to marking at the beginning of sessions"*
>
> (Subject leader interview, November 2013).

Evidence of both teacher marking and pupil responses was seen in children's science books. Some of the teacher marking included numerical scores, which Butler (1988) had found

cancelled out the positive effect of feedback via comment-only marking. Black and Harrison (2010) also argue that the score signifies that the process is complete; the judgement has already been made. They recommend that comment-only marking is used, with any scores recorded for the teacher tracking only.

Summative teacher assessment

When asked how she made a summative judgement, the Year 6 teacher replied:

> "*It's best fit, look at child's work over term, teacher judgement about where work fits and give sublevel. Sometimes do end of term something which can be part of information, but does not 'give' you a level, it informs. There is no set model"*
>
> (Year 6 teacher interview, January 2014).

The *'best fit'* teacher assessment is described as drawing on a range of information which may include a *'child's work'* in normal lessons or an end of term task or question. The Y6 teacher emphasised that the *'end of term something'* does not *'give you a level'*. This is perhaps highlighting the difference between end of Key Stage assessment procedures for different subjects, for example, when the pupils sit a reading test and the score would be converted, by a pre-defined formula, into a level: the test would 'give' the level. In contrast, for a teacher assessment in science, there is no calculation or pre-defined formula, *'no set model'*, to provide a 'best fit' judgement.

For 'best fit' summative assessment, the teacher aims to find the closest match between pupil outcomes and National Curriculum criteria. Such an assessment could enhance validity by reducing the construct under-representation inherent in testing (Gardner et al., 2010) and enhance reliability since teacher assessment can utilise more evidence than is available through external assessment instruments (Mansell, James & the Assessment Reform Group, 2009). However, the lack of transparent processes for collating a term's work into a summative judgement, both opens teacher assessment up to criticisms of bias, especially if the judgements form part of the school's accountability measures (Green & Oates, 2009), together with making it very difficult to explain the processes to others in the community. It also requires a large amount of knowledge of the subject on the part of the teacher, the teacher being entirely responsible for judging whether the pupil's answers are consistent with the teacher's 'model' or expectation of how the pupil can demonstrate understanding (Black & Wiliam, 1998). Connelly, Klenowski and Wyatt-Smith (2012) note that teacher judgements do more than match evidence to criteria, they draw on multiple sources of knowledge, of pupils and previous experience. Without guidance and exemplification, an inexperienced teacher may struggle to make a 'best fit' teacher assessment because they lack a clear expectation of what it would look like for pupils to demonstrate understanding in a topic, and there is a lack of transparent processes for combining such assessments into a 'best fit' judgement.

In addition to concerns regarding the amount of subject-specific knowledge needed to make 'best fit' judgements, another criticism of such an approach is the way that a 'best fit' model produces an overall judgement which could mask gaps in understanding. This was one of the reasons behind the removal of levels in the English National Curriculum, changing to an 'age-related expectations' model whereby pupils would need to meet all criteria (Department for

Education, 2013). The language of summative assessment at School A followed the statutory change from 'levels' at the beginning of the case study period, to 'meeting expectations' by the end of the case study period, but the process for making the summative assessment judgements appeared to continue to be one of 'best fit'. The school's progression grids which were used as criterion scales, were also used throughout the case study period, with only minor alterations to remove the levelling vocabulary for scientific inquiry. Thus it appeared that school processes were resistant to change during the case study period.

The relationship between formative and summative assessment

The science subject leader at School A defined the key purpose of assessment as formative, as Assessment for Learning (AfL):

> *"The purpose of assessment is to develop learning, to identify where children are, and to plan next steps. Assessment should involve children (AfL) and include some success criteria. It should also involve listening and questioning."*
>
> (Subject leader interview, November 2013)

She goes on to state that this formative classroom assessment is utilized when making summative judgements:

> "*The summative judgement arises from formative assessment… a whole school decision was made that summative would be informed by formative leading to a best fit model."*
>
> (Subject leader interview, November 2013)

The use of such a 'formative to summative' model was the reason for choosing this school as a case study, to explore such practice in action, but as noted above, it appears that the notion of 'best fit' summative judgements require a lot of knowledge on the part of the teacher. Underlying the 'formative to summative' model represented by Nuffield (2012) and TAPS (Earle et al., 2016) and enacted in School A, is a shared understanding of progression in science, with explicit criteria or curricular expectations which are, for example, recorded in planning and shared in lessons.

In order to build such a shared understanding of progression and criterion-referenced assessment, a key feature of practice at School A was the allocation of staff development time to science. During regular whole school staff meetings the subject leader introduced new strategies for formative assessment and led the staff in moderation discussions to support summative judgements. As Deputy Head, she would also support staff with planning and teaching using the school's planning and criteria structures.

Harlen (2007) argues that whilst teacher assessment is often perceived as having low reliability, with effective moderation procedures, the reliability of teacher assessment can be as high as it needs to be, in the 'trade off' between reliability and validity (Wiliam 2003). School A appear to be using moderation staff meeting discussions to serve multiple purposes, more than a checking of judgements, it was also a means of staff development (Green & Oates, 2009), supporting both teacher 'assessment literacy' and teacher understanding of progression in science.

DISCUSSION AND CONCLUSIONS

Active pupil involvement in assessment during lessons became the base layer of the TAPS pyramid model (Earle et al., 2016), with many of the examples being provided by School A, including those for self and peer assessment. The school's structures, for example, their progression grids for scientific inquiry, provided explicit criteria for both pupils and teachers to use in their judgements, which together with the moderation discussions, appeared to build a shared understanding of progression in science across the school. This shared understanding was found to be a key feature in other PSTT award-winning schools (Earle, 2015) and became a central addition to the TAPS pyramid model at the monitoring layer (Davies et al., 2017).

The explicit criteria within the school planning structures were used to support both formative and summative assessment, providing success criteria within the lesson and criteria for summative judgements. It is perhaps these common criteria which provide a bridge between formative and summative assessment, providing opportunities for the same classroom activities to be used to inform both formative next steps and summaries of learning.

The case study of School A supported the development of the TAPS pyramid model, both in the addition of new criteria and the provision of exemplification materials. Nevertheless, many of its structures were based on a previous version of the English National Curriculum, when summative assessments used a system of levelling. Unlike other TAPS project schools, little change was seen in assessment practices, perhaps suggesting a one-sidedness in the DBR collaboration. Perhaps School A and its subject leader viewed their role as a provider of examples, rather than as a co-researcher, since they felt that they had already found a way to use formative assessment to inform summative judgements. The lack of change over time in School A is impossible to reduce to the influence of one factor, but recognition of the potential effect of stagnation from over-exemplification is useful to be aware of for future iterations of DBR.

The aim of the TAPS pyramid model is to present assessment principles, supported by a range of exemplification from different contexts, to enable the user to self-evaluate their individual context. This case study of assessment practice at School A has provided some examples for the use of formative assessment in primary science, and the way this information can be summarized to provide a summative judgement.

However, questions have been raised about the school's use of 'best fit' and it is not assumed that this is the only way to implement a 'formative to summative' model of teacher assessment. Additional research is needed to explore practice in other schools and contexts to further test, develop and exemplify the model of teacher assessment, an ongoing focus for the next phases of the TAPS project.

ACKNOWLEDGEMENT

The work presented here is part of the Teacher Assessment in Primary Science (TAPS) project is based at Bath Spa University and funded by the Primary Science Teaching Trust:
https://pstt.org.uk/resources/curriculum-materials/assessment

REFERENCES

Alexander, R. (2008). *Towards Dialogic Teaching – rethinking classroom talk.* Cambridge: Dialogos.

Anderson, T., & Shattuck, J. (2012). Design-based research: a decade of progress in education research? *Educational Researcher*, *41*(1), 16-25.

Black, P., & Harrison, C. (2010). Formative assessment in science. In J. Osborne and J. Dillon (Eds) *Good practice in science teaching: what research has to say*. Maidenhead: Open University Press.

Black, P., & Wiliam, D. (1998). *Inside the black box.* London: GL Assessment.

Bryman (2016). *Social Research Methods 5th Edition*, Maidenhead: OUP.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology*, *58*, 1-14.

Connolly, S., Klenowski, V. & Wyatt-Smith, C. (2012). Moderation and consistency of teacher judgement: teachers' views. *British Educational Research Journal*, *38*, 4, 593-614.

Davies, D., Earle, S., Collier, C., Digby, R., Howe, A., & McMahon, K. (2016). 'Teacher assessment of science in English primary schools.' In: Lavonen, J, Juuti, K, Lampiselkä, J, Uitto, A and Hahl, K, eds. *Electronic proceedings of the ESERA 2015 Conference - Science Education Research: Engaging Learners for a Sustainable Future*. (co-ed. Jens Dolin & Per Kind) (Pt. 11). University of Helsinki, Helsinki, pp. 1577-1588.

Davies, D., Earle, S., McMahon, K., Howe, A., & Collier, C. (2017). Development and exemplification of a model for Teacher Assessment in Primary Science. *International Journal of Science Education*, *39*(14), 1869-1890.

Department for Education (DfE) (2013). *National Curriculum in England: science programmes of study*. London: DfE.

Earle, S. (2014). Formative and summative assessment of science in English primary schools: evidence from the Primary Science Quality Mark. *Research in Science and Technological Education*, *32*(2), 216-228.

Earle, S. (2015). An exploration of whole school assessment systems. *Primary Science 136*, 20-22.

Earle, S., McMahon, K., Collier, C., Howe, A. & Davies, D. (2016). *The Teacher Assessment in Primary Science (TAPS) school self-evaluation tool.* Bristol: Primary Science Teaching Trust.

Gardner, J., Harlen, W., Hayward, L., Stobart, G. with Montgomery, M. (2010). *Developing teacher assessment.* Maidenhead: OUP.

Green, S., & Oates, T. (2009). Considering the alternatives to national assessment arrangements in England: possibilities and opportunities. *Educational Research*, *51*(2), 229-245.

Harlen, W. (2007). *Assessment of learning*. London: Sage.

Harlen, W. (2013). *Assessment and inquiry-based science education: Issues in policy and practice*. Trieste: Global Network of Science Academies.

Hodgson, C., & Pyle, K. (2010). *A Literature Review of Assessment for Learning in Science*. Slough: Nfer.

Mansell, W., James, M., & the Assessment Reform Group (2009). *Assessment in schools: fit for purpose?* London: Teaching and Learning Research Programme.

Mortimer, E., & Scott, P. (2003). *Meaning making in secondary science classrooms*. Maidenhead: Open University Press.

Nuffield Foundation (2012). *Developing policy, principles and practice in primary school science assessment.* London: Nuffield Foundation.

Stake, R. (2006). *Multiple Case Study Analysis*. New York: Guilford Press.

Stobart, G. (2008). *Testing Times: The Uses and Abuses of Assessment*, London: Routledge.

Wiliam, D. (2003). National curriculum assessment: how to make it better. *Research Papers in Education*, *18*, 2, 129–136.

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington: Solution Tree Press.

Wiliam, D., & Black, P. (1996). Meanings and Consequences: A Basis for Distinguishing Formative and Summative Functions of Assessment? *British Educational Research Journal*, *22*, 5, 537-548.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

COMPLEXITY OF PRACTICAL WORK IN SCIENCE CURRICULA AND NATIONAL EXAMS: ANALYSIS OF RECONTEXTUALISING PROCESSES

Sílvia Ferreira and Ana M. Morais
UIDEF, Instituto de Educação, Universidade de Lisboa, Portugal

The study is focused on the level of complexity of practical work in science curricula and national external assessment with regard to the secondary school discipline of Biology and Geology in Portugal. This level of complexity is appreciated through the conceptual demand of practical work as given by the complexity of scientific knowledge and cognitive skills and the relation between theory and practice. The recontextualising processes that may have occurred in the exams were analysed by studying the relation between curriculum and exams. The study makes use of theories and concepts of the areas of psychology and sociology, particularly Bernstein's theory of pedagogic discourse. The results show that the level of conceptual demand of practical work varies according to the specific curricular text under analysis, i.e. Biology or Geology. Practical work as assessed in the exams recontextualises the curriculum in the direction of lowering its level of conceptual demand. In methodological terms, the article explores assumptions used in the analysis and presents innovative instruments.

Keywords: practical work; science process skills; conceptual demand

INTRODUCTION

The role of practical work in offering students the opportunity to experience the process of scientific investigation is one of the arguments for practical work in science education (Hofstein & Kind, 2012; Lunetta et al., 2007; Osborne, 2015). Students are expected to both learn scientific knowledge and mobilize science process skills whenever they are doing investigative practical activities. The nature and complexity of practical work in science curricula and national exams and the recontextualising processes that may have occurred between them should be analysed and discussed because these are aspects that broadly guide textbook authors and teachers' practices.

In science education, as well in other areas of knowledge, it is essential that there are no discontinuities between curriculum, pedagogical practice and assessment (e.g. Britton & Schneider, 2007; Duschl, Schweingruber & Shouse, 2007). For that reason these different texts and contexts "should be conceived of, designed, and implemented as a coordinated system" (Duschl et al., 2007, p. 347). In the specific case of external assessment, evidence from several studies indicates that national exams limit the teaching and learning process and also the classroom assessment tools (Hamilton, 2003). If exams and curriculum are inconsistent, teachers tend to focus on what is assessed in the exams rather than on what is presented in the curriculum and in this way the content that is not tested tends to be ignored in pedagogical practice (Britton & Schneider, 2007). The external assessment can push "teaching and learning in undesirable directions that are counterproductive to the goals of scientific literacy" (p. 1009). However specific types of assessment have the potential to promote particular forms of effective teaching.

The study is focused on the analysis of both the Portuguese curriculum and the national exams for secondary school biannual discipline of Biology and Geology (ages 16^{-}-17^{+}). In Portugal likewise many Latin countries, Biology and Geology, although epistemologically distinct, have traditionally been part of the same discipline (often but not always called Natural Sciences). Theoretically, the study is multidisciplinary, making use of theories and concepts of the areas of psychology and sociology, particularly Bernstein's theory of pedagogic discourse (1990, 2000).

Bernstein develops a theory about the production and reproduction of pedagogic discourse, in which he considers the complex set of relations between various fields and contexts of what he calls pedagogic device. Throughout this process, recontextualisations at the various levels of the pedagogic device can take place and for that reason the pedagogic discourse is not the mechanical result of the dominant principles of society, which constitute the general regulative discourse (GRD). As a result of the official recontextualisation of the GRD, namely at the level of the Ministry of Education and its agencies, the official pedagogic discourse (OPD) is produced. This discourse is expressed, for example, in curricula and in national exams.

Bernstein's model also evidences that the official recontextualisation field is influenced by the fields of economy and symbolic control and defines *the what* and *the how* of the pedagogic discourse. *The what* refers to the knowledge and skills that are the object of the teaching and learning process and *the how* is related to the way in which the teaching and learning process occurs.

In particular the relation between curricula and national exams was analysed in this study to explore recontextualisation processes that may have occurred between the message conveyed in these official documents, with regard to different dimensions of *the what* and *the how* of pedagogic discourse related to practical work. The study addresses the following research problem: What are the messages transmitted by the official pedagogic discourse (OPD) expressed in both the curriculum and the national exams of Biology and Geology of secondary school, with regard to their level of complexity of practical work, and what is the extent to which recontextualising processes do occur?

Varying with authors, practical work can have different meanings. Hodson (1993) considers practical work as a broad concept which includes any activity that requires students to be active. Millar, Maréchal e Tiberghien (1999) limit the definition presented by Hodson (1993) to consider that practical work is 'all those kinds of learning activities in science which involve students at some point handling or observing real objects or materials (or direct representations of these, in a simulation or video-recording)' (p. 36). In the same line, Lunetta, Hofstein and Clough (2007) give the following definition of practical work: 'learning experiences in which students interact with materials or with secondary sources of data to observe and understand the natural world' (p. 394).

The meaning of practical work in the present study is made more precise in that considers that it must mobilize science processes skills. These skills were considered as ways of thinking more directly involved in scientific research, such as observing, formulating problems and hypotheses, controlling variables and predicting (Duschl, Schweingruber and Shouse, 2007). Thus, practical work is defined as: all teaching and learning activities in the sciences in which

the student is actively involved and that allow the mobilization of science processes skills and scientific knowledge and that may be materialized by paper and pencil activities or observing and/or manipulating materials.

The level of complexity of practical work can be appreciated by its level of conceptual demand. In the context of the research that has been carried out by the ESSA Group (Sociological Studies in the Classroom, Institute of Education, University of Lisbon) within Bernstein's theory, the concept of conceptual demand is defined as the level of complexity of science education as given by the complexity of scientific knowledge and the strength of intradisciplinary relations between distinct knowledge and also by the complexity of cognitive skills (Morais & Neves, 2016).

METHOD

The analysis of the Biology and Geology secondary school curriculum was focused on two official documents which contain directions for the teacher: 10th Biology and Geology syllabus and 11th Biology and Geology syllabus (in force since 2002 and 2003, respectively). Although part of the same discipline and of the same curriculum, Biology and Geology come in the curriculum as two distinct subjects, with strong boundaries between them. The analysis of the national exams involved 26 exams, from 2006 to 2011.

The whole curriculum was segmented into units of analysis but the units of analysis with a specific reference to practical work (requiring the mobilization of science process skills) were the only ones considered in this study. For the same reason, the analysis of national exams considered only the questions which focused on practical work, i.e., questions that mobilised science process skills. Each question was taken as a unit of analysis.

The level of conceptual demand was determined through the analysis of specific dimensions of *the what* and of *the how* of the OPD (Figure 1). The first corresponds to the level of complexity of scientific knowledge and cognitive skills and the second to the strength of intradisciplinary relations between theory and practice. The discontinuities between the curriculum and the national exams were studied through the recontextualising processes that may have occurred between the messages of these official documents.

Three instruments were constructed in order to characterise the message underlying each one of the units of analysis, and consequently the OPD transmitted by both the science curriculum and the national exams, with regard to the conceptual demand of practical work. The construction of the instruments followed a mixed methodology (Creswell & Clark, 2011; Morais & Neves, 2010; Teddlie & Tashakkori, 2009), using qualitative and quantitative approaches. The text that follows contains a brief description of the instruments constructed and how they were used, and gives also some examples to show how analyses were made.

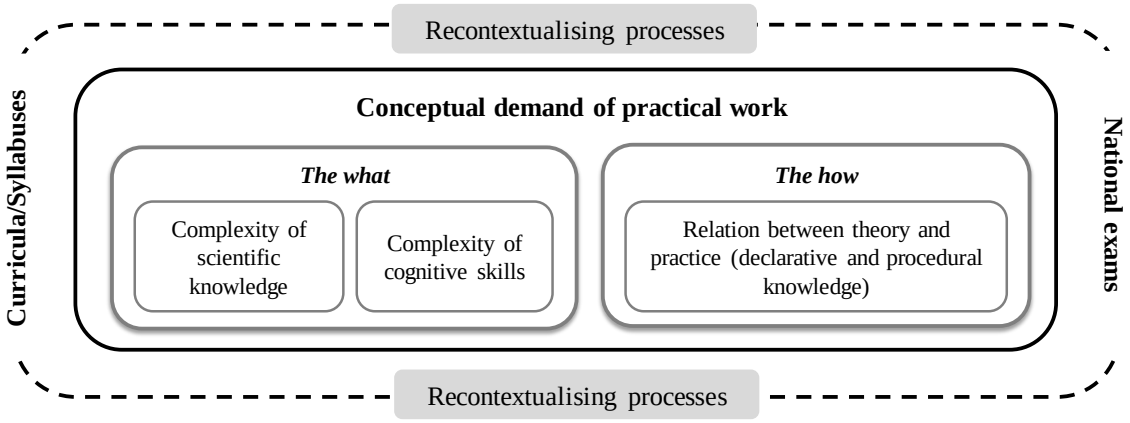


Figure 1. Diagram of dimensions, related to *the what* and *the how* of practical work, analysed in the secondary school Biology and Geology curriculum and national exams.

The instrument for analysing the complexity of scientific knowledge was based on the distinction between facts, generalized facts, simple concepts, complex concepts and unifying themes/theories, following several authors (e.g. Brandwein et al., 1980; Cantu & Herron, 1978; Duschl et al., 2007; Pella & Voelker, 1968). For instance, simple concepts have a low level of abstraction, defining attributes and examples that are observable (Cantu & Herron, 1978). Complex concepts correspond to abstract concepts proposed by Cantu and Herron (1978) and "are those that do not have perceptible instances or have relevant or defining attributes that are not perceptible" (p.135). Table 1 presents an excerpt of this instrument and examples of units of analysis which illustrate different degrees of complexity.

Table 1. Excerpt of the instrument to characterise the complexity of scientific knowledge.

Degree 1	Degree 2	Degree 3	Degree 4
Scientific knowledge of low level of complexity, as facts, is referred.	Scientific knowledge of level of complexity greater than degree 1, as simple concepts, is referred.	Scientific knowledge of level of complexity greater than degree 2, as complex concepts, is referred.	Scientific knowledge of very high level of complexity, as unifying themes and theories, is referred.

Units of analysis:

Degree 1: [1] Search for information on the internet, in newspapers and magazines about the consequences of such situations [anthropic occupation of floodplains and coastal zones, and construction in slope zones] for populations. (*11th Geology syllabus)*.

Degree 2: [2] [...] 6. When exposed to the sun, the surface of the coat of *C. dromedarius* can reach temperatures above 70 °C, while at the skin level the body temperature does not exceed 40 °C. Explain, from the data provided, how the research carried out allowed to relate the adaptation to high temperatures to the levels of transpiration presented by *C. dromedarius*. [...] (*National Exam of 2009, 1st phase*)

Degree 3: [3] [...] 6. Genetic studies in *Coccomyxa* suggest that as soon as the endosymbiotic relation with *Ginkgo biloba* was established the algae was transmitted from generation to generation. Explain how the results of those studies may relate the transmission of the endosymbiotic relation from generation to generation to the way how such relation was initiated. [...] (*National Exam of 2009, 2nd phase*).

Degree 4: [4] Collect, organize and interpret data of a different nature related to evolutionism and to arguments that support it by opposition to fixism. (*11th Biology syllabus*).

Adapted from Ferreira & Morais (2013, 2014)

Excerpt [1] emphasises facts related to the consequences for populations of the anthropic occupation of floodplains and coastal zones and construction in slope zones, and for that reason it was classified with the degree 1. In excerpt [2], the national exam question and respective recommended correction involve simple concepts related to thermoregulation. In the question presented in excerpt [3] and in the given respective correction are involved complex concepts related to the genetic transmission of an endosymbiotic relation between a plant and a green algae. If the question appealed to a relation to the model of endosymbiosis, the degree of complexity would increase to degree 4. The excerpt [4] focuses knowledge of a very high degree of complexity related to the theory of evolution.

A second instrument, for analysing the complexity of cognitive skills, was based on the taxonomy created by Marzano and Kendall (2007, 2008) with four levels for the cognitive system: retrieval, comprehension, analysis and knowledge utilization. Retrieval, the first level of the cognitive system, involves the activation and transfer of knowledge from permanent memory to working memory. "The process of comprehension within the cognitive system is responsible for translating knowledge into a form appropriate for storage in permanent memory" (2007, p.40). The third level, analysis, involves the production of new information that the individual can elaborate on the basis of the knowledge s/he has comprehended. The fourth and more complex level of the cognitive system implies the knowledge utilization in concrete situations. Table 2 presents an excerpt of this instrument.

Table 2. Excerpt of the instrument to characterise the complexity of cognitive skills.

Degree 1	Degree 2	Degree 3	Degree 4
Cognitive skills of low level of complexity, involving cognitive processes of retrieval, are mentioned.	Cognitive skills of level of complexity greater than degree 1, involving cognitive processes of comprehension, are mentioned.	Cognitive skills of level of complexity greater than degree 2, involving cognitive processes of analysis, are mentioned.	Cognitive skills of very high level of complexity, involving cognitive processes of knowledge utilization, are mentioned.
Units of analysis:			
Degree 1: *No units of analysis were found.*			
Degree 2: [5] [...] 3.2. Select the alternative that completes the following statement correctly. For the results of Büchner's experiment prove that the occurrence of fermentation is in some way related to the intervention of living beings (or their derivatives), it would be necessary to introduce in the procedure a device containing ... (A) ... yeast in a sugar solution. (B) ... yeast extract in a sugar solution. (C) ... only a sugar solution. (D) ... exclusively yeasts. (*National Exam of 2007, 2nd phase*)			
Degree 3: [6] Classify rocks based on genetic and textural criteria. (*11th Geology syllabus*)			
Degree 4: [7] [...] 6. Some authors consider Giardia as a missing link in the evolution between prokaryotic cells and eukaryotic cells, while others authors argue that it has evolved from more complex eukaryotic cells by the loss of certain organelles. Present a possible path of investigation that would allow one of the hypotheses mentioned to be proved and the other to be rejected. [...] (*National Exam of 2006, 1st phase*)			

Adapted from Ferreira & Morais (2013, 2014)

In excerpt [5] the national exam question implies the mobilization of science process skills related to the identification of the control group characteristics, which is associated with the

process of comprehension. The syllabus aim presented in excerpt [6] involves the mental process of classification, associated with the cognitive process of analysis. The excerpt [7] focuses the planning of investigative laboratory activities, which is related to the cognitive process of knowledge utilization.

The analysis of the relation between theory and practice was characterized through Bernstein's concept of classification (1990, 2000), to indicate the strength of boundaries between various types of knowledge. This instrument contained a four degree scale of classification (C^{--}, C^{-}, C^{+}, C^{++}). The weakest classification (C^{--}) corresponds to an integration of theory and practice where both have equal status and the strongest classification (C^{++}) corresponds to an insulation between theory and practice. The empirical descriptors for each degree translate the relation between declarative knowledge (theory) and procedural knowledge (practice) (Roberts, Gott & Glaesser, 2010). Table 3 presents an excerpt of this instrument, followed by examples of units of analysis which illustrate different levels of classification.

Table 3. Excerpt of the instrument to characterise the relation between theory (declarative knowledge) and practice (procedural knowledge).

C^{++}	C^{+}	C^{-}	C^{--}
The focus is either on declarative knowledge only or on procedural knowledge only.	Declarative knowledge and procedural knowledge are focused, but not the relation between them.	The relation between declarative and procedural knowledge is focused, giving higher status to declarative knowledge.	The relation between declarative and procedural knowledge is focused, giving equal status to both types of knowledge.

Units of analysis:

C^{++}: [8] [...] 3. Select the alternative that fills the spaces in the following sentence, in order to get a correct statement. The study II allows to conclude, through the quantification of the seeds produced, that the _____ space selected plants with _____ dispersion capacity.
(A) urban (...) greater
(B) country (...) greater
(C) urban (...) minor
(D) country (...) minor (*National Exam of 2008, 1st phase*)

C^{+}: *No units of analysis were found.*

C^{-}: [9] The cell: The laboratory observation of uni and multicellular living beings, collected in the field, will enable the understanding of the cell as a structural and functional unit of living beings and facilitate the approach to its basic constituents. (*10th Biology syllabus*)

C^{--}: [10] Create models and simulate laboratory situations of landslide, trying to identify the factors that contribute to their occurrence. The teacher should draw attention to the analogies between the model and the geological process, stressing, however, the variables involved and the different scales of time and space in which phenomena occur. (*10th Geology syllabus*)

Adapted from Ferreira & Morais (2013, 2014)

The national exam question presented in excerpt [8] focuses on procedural knowledge only, associated with the knowledge of the scientific process of interpretation of simple experimental data, explored in the introductory text of this question. The excerpts [9] and [10] involve a relation between declarative and procedural knowledge, but in the former the higher status is given to declarative knowledge about the cell, and in the latter both types of knowledge have equal status.

In order to clarify how the same unit of analysis was classified in the study in terms of the dimensions related to *the what* and *the how* of pedagogic discourse, an illustrative example of the analysis that was made is presented:

> [11] Setting experimental devices with simple aerobic facultative living beings (e.g. *Saccharomyces cerevisae*) in nutritive media (e.g. "bread dough", grape juice, aqueous solution of glucose…) with different degrees of aerobiosis. Identification with the students of the variables to be controlled and the indicators of the process under study (e.g. presence/ absence of ethanol). (*10th Biology syllabus*)

Excerpt [11] presents a methodological guideline of the 10th Biology syllabus. With regard to *the what* of the OPD, this unit is focused on a laboratory activity, which appeals to simple concepts, related to glucose degradation in the presence and in the absence of oxygen (degree 2), and to cognitive skills involving the cognitive process of analysis, since it implicates the control of variables (degree 3). With regard to *the how* of the OPD, this unit of analysis involves a relation between declarative and procedural scientific knowledge, where equal status is given to these two types of knowledge (C^{--}).

RESULTS

Figure 2 gives a synthesis of results of the conceptual demand of practical work of both science curriculum and national exams for the three dimensions studied. These results refer to the Biology and Geology curriculum specific guidelines only and to the national exams from 2006 to 2011.

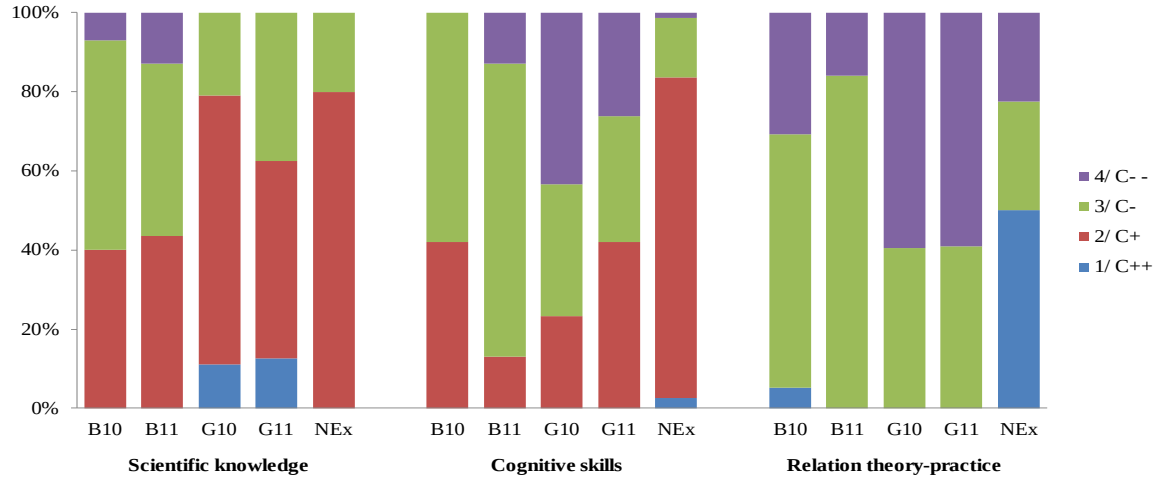


Figure 2. Conceptual demand of practical work in Portuguese Biology and Geology curriculum and external assessment at secondary school (*B10* 10th Biology syllabus, *B11* 11th Biology syllabus, *G10* 10th Geology syllabus, *G11* 11th Geology syllabus, *NEx* National Exams).

When Biology and Geology curricular subjects are compared, Biology shows more complex concepts and unifying themes (degrees 3 and 4) than Geology. The higher knowledge complexity in Biology practical work is especially given by the focus on cell theory and on evolution theory. In the case of Geology there are no units classified with degree 4 and there are units classified with degree 1. Simple concepts prevail in exams (degree 2). Degrees 1 and 4 (facts and unifying themes/theories, respectively) are absent in exams questions about practical work.

When the focus is the complexity of cognitive skills, it is Geology that places greater emphasis on complex cognitive skills of a high level (cognitive process of knowledge utilization – degree 4) when compared with Biology. The highest complexity of cognitive skills in Geology practical work is particularly related to the formulation of hypotheses, decision making, construction of models and research, organization and processing of information. Exams questions that mobilised science process skills were focused on the cognitive process of comprehension (degree 2).

With regard to the relation between theory and practice, most units were classified with C^{-} in Biology which correspond to the units that reflect a relation between the two types of knowledge with a focus on declarative knowledge. The data of Figure 2 also shows that C^{--} prevails in Geology syllabus which means that most units suggest a relation between declarative and procedural scientific knowledge, equal status being given to these two types of knowledge. In the exams half of the questions were classified with C^{++}. This classification refers to the second part of the respective instrument descriptor (Table 3), that is these questions only present procedural knowledge.

DISCUSSION AND CONCLUSIONS

The present study intended to appreciate the recontextualising processes that may have occurred between the messages expressed in the curriculum and the national exams of the Biology and Geology discipline in relation to the complexity of practical work. The results show the occurrence of discontinuities between the messages of the curriculum and the external assessment. Although the analysis is focused on the Portuguese educational system, the findings and methodologies of this study may be extended to other studies and may give a contribution to raising the level of conceptual demand of practical work in science education.

Through the analysis of the complexity of scientific knowledge and cognitive skills and the relation between theory and practice, it was possible to appreciate the level of conceptual demand of practical work expressed in the Official Pedagogic Discourse. When the discipline is taken as a whole (Biology and Geology together), the results evidence a considerable level of conceptual demand of practical work. However the separate analysis of the two subjects shows that Biology has a generally higher level of conceptual demand when compared with Geology. Practical work assessment in the national exams has a low level of conceptual demand, showing recontextualisation processes in the direction of lowering the level of the curriculum.

Within the curriculum have also occurred recontextualisation processes between the messages of practical work in Biology and Geology, considered as two separate components of the same discipline. One possible explanation for these discontinuities is related to the Ministry of Education selection of two different teams of authors to construct the curriculum of each one of the curricular areas. Each team of authors seemed to value different dimensions of *the what* and *the how* of pedagogic discourse. Some of these differences may also be related to the fact that Biology and Geology, although in Portugal are part of the same discipline, are epistemologically distinct curricular areas. In the case of the external assessment, the level of conceptual demand of practical work is lower than the level of the curriculum, namely in the

case of the Biology syllabuses (the area most valued in the exams questions about practical work).

With regard to the complexity of scientific knowledge, the external assessment of practical work mainly values simple concepts. There is therefore a discontinuity between assessment and the curriculum practical work messages, where the Biology syllabuses give more emphasis to complex scientific knowledge (complex concepts and unifying themes/ theories). If science education is to reflect the structure of scientific knowledge then it should lead to the understanding of concepts and big ideas, although that understanding requires a balance between knowledge of distinct levels of complexity (Morais & Neves, 2016). Bybee and Scotter (2007) also present this aspect as a principle for the development of an effective science curriculum.

When the focus is the complexity of cognitive skills, the external assessment gives greater emphasis to simple skills, especially those involving the cognitive processes of comprehension. Similarly to scientific knowledge, in this case there is also a discontinuity in relation to the message of the Biology syllabuses in which complex skills prevail, particularly those associated with the cognitive process of analysis. The situation that better represents an efficient scientific learning, when practical work is implemented, is a situation where there is a balance between complex and simple cognitive skills. In this way, only when students develop simple skills, such as the memorization of certain facts and concepts, can they develop complex skills, such as applying these concepts to new situations (Geake, 2009).

In the case of the relation between theory and practice, there is also a devaluing of this relation when passing from the Biology and Geology curriculum to the national exams. For example while in the Biology syllabuses there is a relation between theory and practice, in the external assessment half of the practical work questions only focused procedural knowledge without relating it to declarative knowledge. The results of external assessment reinforces the results of other studies (e.g., Abrahams & Millar, 2008) that point out to the existence of a separation between theory and practice when teachers implement practical activities, particularly laboratory work.

In this study it was considered that the desirable situation with respect to the relation between theory and practice is a situation in which relations between declarative and procedural knowledge predominate, with more status being given to declarative knowledge in the relation. This is the situation that best represents an efficient scientific learning that is learning that is supported by the understanding and applying of science processes knowledge. The Biology syllabuses are closer to that situation.

The results of this study show that the external assessment presents a low level of conceptual demand, evidencing recontextualisation processes that reduce the level of the Biology and Geology curriculum. These are results of particular concern because external assessment tends globally to influence the curriculum in practice and specifically to condition textbook authors and teachers' practices. All knowledge and skills that are not the subject of external assessment tends to be ignored in pedagogic practice (e.g., Britton & Schneider, 2007).

The study highlights a major issue of educational systems that are not horizontally coherent i.e. systems where assessment is not aligned with the curriculum. As Wilson and Bertenthal (2006) refer, "to serve its function well, assessment must be tightly linked to curriculum and instruction so that all three elements are directed toward the same goals" (p. 4).

ACKNOWLEDGEMENT

The authors acknowledge to Isabel Neves for her contribution in the analysis of the curriculum and national exams and in the manuscript revision. This research was partially financed by the Foundation for Science and Technology (SFRH/BD/68346/2010).

REFERENCES

Abrahams, I., & Millar, R. (2008). Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *International Journal of Science Education, 30*(14), 1945-1969.

Bernstein, B. (1990). *Class, codes and control: Volume IV, The structuring of pedagogic discourse.* London: Routledge.

Bernstein, B. (2000). *Pedagogy, symbolic control and identity: Theory, research, critique (rev. ed.).* Londres: Rowman & Littlefield.

Brandwein, P., Cooper, E., Blackwood, P., Cottom-Winslow, M., Boeschen, J., Giddings, M., Romero, F., & Carin, A. (1980). *Concepts in science – Teacher's edition*. New York: Harcourt Brace Jovanovich.

Britton, E. D., & Schneider, S. A. (2007). Large-scale assessments in science education. In N. Lederman & S. Abel (Eds,), *Handbook of research on science education* (pp.1007-1040). Mahwah, NJ: Lawrence Erlbaum.

Bybee, R. W., & Scotter, P. (2007). Reinventing the science curriculum. *Educational Leadership, 64*(4), 43-47.

Cantu, L. L., & Herron, J. D. (1978). Concrete and formal Piagetian stages and science concept attainment. *Journal of Research in Science Teaching, 15*(2), 135-143.

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conduction mixed methods research* (2.ª ed.). Thousand Oaks, CA: Sage.

Duschl, R., Schweingruber, H., & Shouse, A. (Ed.) (2007). *Taking science to school: Learning and teaching science in grade K-8*. Washington: National Academies Press.

Ferreira, S., & Morais, A. (2013). Exigência conceptual do trabalho prático nos exames nacionais: Uma abordagem metodológica. *Olhar de Professor, 16*(1), 149-172.

Ferreira, S., & Morais, A. M. (2014). Conceptual demand of practical work in science curricula: A methodological approach. *Research in Science Education, 44*(1), 53-80.

Geake, J. (2009). *The brain at school: Educational neuroscience in the classroom*. Berkshire, UK: Open University Press.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education, 27*, 25-68.

Hodson D (1993). Re-thinking old ways: Towards a more critical approach to practical work in school science. *Studies in Science Education, 22*(1), 85-142.

Hofstein, A., & Kind, P. M. (2012). Learning in and from science laboratories. In J. Fraser, K. Tobin & C. J. McRobbie (eds.), *Second International Handbook of Science Education* (pp.189-207). New York: Springer.

Lunetta, V. N., Hofstein, A. & Clough, M. (2007). Learning and teaching in the school science laboratory: An analysis of research, theory, and practice. In N. Lederman & S. Abel (Eds,), *Handbook of research on science education* (pp.393-441). Mahwah, NJ: Lawrence Erlbaum.

Marzano, R. J., & Kendall, J. S. (2007). *The new taxonomy of educational objectives* (2.ª ed.). Thousand Oaks, CA: Corwin Press.

Marzano, R. J., & Kendall, J. S. (2008). *Designing & assessing educational objectives: Applying the new taxonomy*. Thousand Oaks, CA: Corwin Press.

Millar, R., Maréchal, J. F., & Tiberghien, A. (1999). Maping the domain – varieties of practical work. In J. Leach & A. Paulsen (Eds.), *Practical work in science education* (pp.33-59). Denmark: Roskilde University Press.

Morais, A. M., & Neves, I. P. (2010). Basil Bernstein as an inspiration for educational research: Specific methodological approaches. In P. Singh, A. Sadovnik & S. Semel (Eds.), *ToolKits, translation devices and conceptual accounts: Essays on Basil Bernstein's sociology of knowledge* (pp. 11-32). New York: Peter Lang.

Morais, A. M., & Neves, I. P. (2016). Vertical discourses and science education: Analyzing conceptual demand of educational texts. In P. Vitale & B. Exley (Eds.), *Pedagogic rights and democratic education: Bernsteinian explorations of curriculum, pedagogy and assessment* (Chap. 13). London: Routledge.

Osborne, J. (2015). Practical work in science: misunderstood and badly used? *School Science Review, 96*(357), 16-24.

Pella, M., & Voelker, A. (1968). Teaching the concepts of physical and chemical change to elementary school children. *Journal of Research in Science Teaching, 5*(4), 311-323.

Roberts, R., Gott, R., & Glaesser, J. (2010). Students' approaches to open-ended science investigation: The importance of substantive and procedural understanding. *Research Papers in Education, 25*(4), 377-407.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: Sage.

Wilson, M., & Bertenthal, M. (Eds.) (2006). *Systems for state science assessment*. Washington, DC: National Academies Press.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

PEER-ASSESSMENT AS A LEARNING ACTIVITY FOR SECONDARY SCHOOL STUDENTS IN MODELING-BASED LEARNING

Olia Tsivitanidou[1], Costas P. Constantinou[1] and Peter Labudde[2]
[1] University of Cyprus, Nicosia, Cyprus
[2] University of Applied Sciences and Arts, Northwestern Switzerland

The aim of this study was to investigate how reciprocal peer assessment in modeling-based learning can serve as a learning activity for secondary school learners in a physics course. The participants were twenty-two upper secondary school students from a Gymnasium in Switzerland. They were asked to model additive and subtractive color mixing in groups of two, after having completed hands-on experiments in the laboratory. Then, they submitted their models and anonymously assessed the model of another peer group. The students were given a 4-point rating scale with pre-specified assessment criteria, while enacting the peer-assessor role. After implementation of the peer assessment, students, as peer assessees, were allowed to revise their models. They were also asked to complete a short questionnaire, reflecting on their revisions. Data were collected by: (i) peer-feedback reports, (ii) students' initial and revised models, (iii) post-instructional interviews with students, and (iv) students' responses to open-ended questions. The data were analyzed qualitatively and then quantitatively. The results revealed that, after enactment of the peer assessment, students' revisions of their models reflected a higher level of attainment toward their model-construction practices and a better conceptual understanding of additive and subtractive color mixing. The findings of this study suggest that reciprocal peer assessment, in which students experience both the role of assessor and assesse, facilitates students' learning in science.

Keywords: reciprocal peer assessment; modeling competence, physics instruction.

INTRODUCTION

Recent developments in the field of assessment stress the importance of formative approaches, in which assessment is realized as part of the learning process to support the improvement of learning outcomes (Bell & Cowie, 2001). Formative assessment has also received emphasis as a mechanism for scaffolding learning in science. Peer assessment, when employed formatively, can improve students' learning accomplishment and their overall performance (e.g., specific skills and practices) in various domains including in Science Education (Grob, 2017; Tsivitanidou & Constantinou, 2016; Tsivitanidou, Zacharias, & Hovardas, 2011; Chen, Wie, Wu & Uden, 2009). Reflection processes can be enhanced in the context of reciprocal peer assessment, in which students can benefit from the enactment of the role of both the assessor and the assessee (Tsivitanidou et al., 2011). Learning gains can emerge when students receive feedback from their peers, but also when they provide feedback to their peers, because they might be introduced to alternative examples and approaches and can also attain significant cognitive progression (Hwang, Hung & Chen, 2014). This renders peer assessment not only an innovative assessment method (Cestone, Levine & Lane, 2008), but also a learning activity (Orsmond, Merry & Reiling, 1996), in the sense of co-construction of knowledge, that is, constructing new knowledge by the exchange of pre-conceptions, questions, and hypotheses (Labudde, 2000). Despite those benefits, few studies have focused on peer assessment in

modeling-based learning (Chang & Chang, 2013). As a result, there is a need to further examine what students are able to do in modeling-based learning, especially in terms of whether the experience of peer assessment could be useful for them and their peers with respect to the enhancement of their learning.

THEORETICAL FRAMEWORK

Reciprocal peer assessment and peer feedback

Peer assessment can be characterized as one-way or two-way / reciprocal / mutual, depending on the particular roles that students enact while implementing it (Hovardas, Tsivitanidou, & Zacharia, 2014). This study focuses on reciprocal peer assessment, that is the type of formative assessment in which students are given the opportunity to assess each other's work, thus enacting both roles of the assessor and the assessee. While enacting the peer-assessor role, students are required to assess peer work and to provide peer feedback for guiding their peers in improving their work (Topping, 2003). In the peer-assessee role, students receive peer feedback and they can further use it for revising their artefacts and ultimately enhance their future learning accomplishments (Tsivitanidou & Constantinou, 2016).

Research findings have shown that, when learners are engaged in both roles of the assessor and the assessee, in the context of reciprocal peer assessment, certain assessment skills are required (Gielen & de Wever, 2015). When enacting the peer-assessor role, students need to be able to assess their peers' work with particular assessment criteria (Sluijmans, 2002), judge the performance of a peer, and eventually provide peer feedback. Apart from assessment skills (Sluijsmans 2002), peer assessment also requires a shared understanding of the learning objectives and content knowledge among students in order to review, clarify, and correct peers' work (Ballantyne, Hughes & Mylonas, 2002). In the role of the peer assessee, students traditionally need to review in a critical manner the peer feedback and decide on whether and how to further utilize it for improving their own work (Hovardas et al., 2014, p. 135). In both cases, reciprocal peer assessment engages students in cognitive activities such as summarizing, explaining, providing feedback and identifying mistakes and gaps, which are dissimilar from the expected performance (Van Lehn, Chi, Baggett, & Murray, 1995).

The provision of peer feedback is also intended to involve students in learning by providing to and receiving from their peers' opinions, ideas, and suggestions for improvement (Hovardas et al., 2014; Black & William, 1998; Kim, 2005). In the context of peer assessment, students receive feedback from peers who share a similar language level/code as their own, which may result in the feedback being more comprehensible (student-speak) compared to a feedback received from the teacher (teacher-speak) or an expert (science-speak). The peers, as assessors, have also had to perform the same task themselves, so might have a good sense of where potential problems/difficulties in executing the task could lie. Their language could speak more directly to the actual features of task performance (than that of an assessor standing outside). In fact, previous studies have revealed that peer feedback might bare more learning benefits to students than expert feedback (Frost & Turner, 2005; Yang, Badger &Yu, 2006).

Although feedback has proven to be advantageous for both learning and performance (e.g., Nelson & Schunn, 2008), it appears that not all types of feedback automatically result in

performance improvement (Kluger & DeNisi, 1996). For example, it has been shown that peer feedback comments including explanatory statements and justification are associated with the effectiveness in enhancing the performance of assessees (Narciss & Huth, 2006). Apart from that, there are certain conditions under which feedback can lead to learning benefits for students. According to Nicol and Macfarlane-Dick (2006) external feedback might be provided to a student by the teacher, by a peer or by other means (e.g. a placement supervisor, a computer). This additional information might augment, concur or conflict with the student's interpretation of the task and the path of learning (Nicol & Macfarlane-Dick, 2006). However, to produce an effect on internal processes or external outcomes the student must actively engage with these external inputs. In effect, any kind of external feedback (provided either by peers or the teacher) has to be interpreted, constructed and internalized by the student if they were to have a significant influence on subsequent learning. Apart from the effect that feedback may entail in students' learning, previous studies, have also revealed that providing feedback may be more beneficial for the assessor's future performance than that of assessees who simply receive feedback (Cho & Cho, 2011; Hwang et al., 2014; Kim, 2009; Nicol, Thomson & Breslin, 2014), since giving feedback is related mainly to critical thinking whereas receiving feedback is related mainly to addressing subject content that needs clarification or other improvement (Hwang et al., 2014; Nicol, et al., 2014). For these reasons, researchers argue that further research on the impact of peer feedback on students' learning and performance is needed (e.g., Tsivitanidou & Constantinou, 2016; Evans, 2013; Hattie & Timperley, 2007).

The modeling competence in science learning

Research focusing on the modeling competence contributed significantly to the overall growth and development of research in science education (Gilbert & Justi, 2016) and that is because scientific models and modeling play an important role in the teaching and learning of science (Acher, Arcà, & Sanmartí, 2007; Hodson, 1993) by introducing learners to scientific ways of reasoning and by linking the worlds of observations and theory (Schwarz, et al., 2009). The modeling competence can be fostered in the context of modeling-based learning (Nicolaou 2010; Papaevripidou 2012), which refers to "learning through construction and refinement of scientific models by students" (Nicolaou, & Constantinou, 2014, p. 55). Papaevripidou, Nicolaou, and Constantinou (2014) proposed the Modeling Competence Framework (MCF) which suggests the breakdown of modeling competence into two categories: *modeling practices* and *meta-knowledge about modeling and models* (Papaevripidou et al., 2014; Nicolaou & Constantinou, 2014). It emerged from a synthesis of the research literature on learning and teaching science through modeling (Papaevripidou et al., 2014). Within this framework, it is suggested that learners' modeling competence emerges as a result of their participation within specific modeling practices, and is shaped by meta-knowledge about models and modeling (Schwarz et al., 2009). In this study, we focused on students' modeling practices of model construction and evaluation, because these are essential processes that lead to successful and complete acquisition of the modeling competence (Chang & Chang, 2013; NRC, 2007; NRC, 2012). In addition, in the context of modeling-based learning, a few studies (e.g., Chang & Chang, 2013; Pluta, Chinn, & Duncan, 2011; Tsivitanidou, Constantinou, Labudde, Rönnebeck, & Ropohl, 2017) have provided evidence specific to the educational

value of teaching-learning activities that involve the evaluation of models by students themselves. The evaluation of models as a process involves engaging students in discussing the quality of models for further improvement and revision (Chang, Quintana, & Krajcik, 2010; Schwarz & Gwekwerere, 2006; Schwarz & White, 2005; Schwarz et al., 2009). Considering that previous research in this direction is scarce (Tsivitanidou, et al., 2017) there is a need to further investigate what students can do when assessing peers' models and how peer assessment, in modeling-based learning, can foster students' model-construction practices, as well as their conceptual understanding of scientific phenomena (Chang & Chang 2013; Pluta et al., 2011).

Objectives of this study

In this study, we aimed to examine whether reciprocal peer assessment, when employed formatively, can facilitate students' learning in science. In particular, we sought to examine how the enactment of the peer-assessor and peer-assessee roles is associated with students' improvements on their own constructed models, after enacting reciprocal peer assessment. In this study, we focused on students' modeling practices of model construction and evaluation (Schwarz & White, 2005). The research question that we sought to address was: *Is there any evidence suggesting that the enactment of reciprocal peer assessment is related to secondary school students' learning benefits in modeling-based learning in the context of Light and Color?*

METHODOLOGY

Participants

The sample consisted by $N = 22$ upper-secondary school students coming from a Gymnasium, in Northwestern Switzerland. Overall, there were almost equal numbers of girls and boys (12 girls and 10 boys). Students worked in randomized pairs in most activities and the pairs remained unchanged throughout the intervention. There were eleven groups of two students (home groups). The home groups were coded with numbers (1 to 11), and within each group, the students were also coded (as Student A and Student B). As confirmed from the post-instructional interviews with eleven participants, most of them ($n = 8$) had experienced oral and / or written peer assessment in the past in different subjects.

Teaching material

The sequence was grounded in collaborative modeling-based learning, during which students were asked to work in their groups and collaboratively construct their model. The students worked through the learning material on the topic of *light and color* in the context of their physics course. The curriculum material required the students to work, in groups of twos (home groups), with a list of hands-on experiments on additive and subtractive color mixing. Those activities lasted four meetings of 45 minutes each (week 1 and 2). In the meeting that followed (week 3), and after having completed the experiments, students in each group were instructed to draw inferences relying on their observations and the gathered data. Their inferences were explicitly expected to lead to a scientific model which can be used to represent, interpret, and predict the additive and subtractive color mixing of light. For doing so, students were provided

with a sheet of paper, color pencils, and a list of specifications that they were asked to consider when developing their model. Finally, in the last meeting (week 4), the students implemented the peer-assessment activity (models exchange, peer review, revision of models). Overall, it took the student groups six meetings (lessons) of 45 minutes to complete this sequence, in a total period of four weeks.

Peer-Assessment Procedure

As soon as the students had finalized their models in their home groups, they exchanged their models with other groups, that is, two groups reciprocally assessed their models (e.g. home group 3 exchanged its model with the model of home group 4). The pairs of groups involved in the exchanges were randomly assigned by the teacher. Peer assessors used a 4-point Likert rating scale with eight pre-specified assessment criteria for accomplishing the assessment task. The assessment criteria were addressing the Representational Power (PP), Interpretive Power (IP) and Predictive Power (PP) of the model and thus they were in line with the list of specifications that was given to the students prior to the model-construction phase. Assessors rated their peers' models on all criteria in accordance with the 4-point Likert scale. Along with the ratings, assessors were instructed to provide assessee groups with written feedback (for each criterion separately), in which they were to explain the reasoning behind their ratings, and provide judgments and suggestions for revisions. On average, it took each peer assessor 15 minutes to complete the assessment (SD = 2.0). Once the students had completed the assessment of their fellow students' models on an individual basis, they provided the feedback that they had produced to the corresponding assessee group. Therefore, each home group received two sets of peer feedback from another peer group. During the revision phase, students in their home groups collaboratively reviewed the two peer-feedback sets received from the corresponding assessor group. Students were free to decide on whether to make any revisions to their model. By the end of the revision phase, students responded, in collaboration with their group mate in their home groups, to two open-ended questions which were given to them for reflection purposes (Question A: *"Did you use your peer's feedback to revise your model? Explain your reasoning."* Question B*: "Did you revise your model after enacting peer assessment? Explain your reasoning."*). By the end of this intervention, eleven students (each from a different home group) were interviewed individually about their experience with the peer-assessment method. Each interview lasted approximately 15 to 20 minutes. Students were first asked about any previous experience in peer assessment; then they were asked whether they found peer assessment, as experienced in this study, useful assuming the role of the peer assessor and peer assessee respectively. Interview and post-instructional questionnaire data were used for triangulation purposes.

Data sources

At the beginning of the intervention, a consent form was signed by the students' parents for allowing us to use the collected data anonymously for research purposes. The following data were collected: (i) students' initial models; (ii) peer-feedback reports; (iii) students' revised models; (iv) post-instructional interviews with eleven students, and (v) home-groups' responses to the two open-ended questions at the end of the intervention.

Data analysis

We used a mixed-methods approach that involved both qualitative and quantitative analyses of the data. In particular, the data were first analyzed qualitatively and then also quantitatively with the use of the SPSS™ software, except for the interviews and students' responses to the two open-ended questions which were only qualitatively analyzed. We examined each student's learning progression as reflected in the quality of their initial and revised models, with respect to the intended learning objectives. In case of revisions applied by students, we further examined possible parameters which might have let the students proceed with the revisions in their models. Inter-rater reliability data were also collected [Krippendorff's Alpha coefficient > 0.79 for the coding of peer-feedback data and initial and revised models; Cohen's Kappa > 0.80 for qualitative (categorical) items].

RESULTS

The data analysis revealed that ten, out of eleven home groups, revised their models, after the enactment of peer assessment. All revisions applied by assessees were found to improve the quality of their initial models; in other words, no case was identified in which assessees proceeded to revisions that undermined the quality of their initial model. This implies that students, as assessees, were able to filter invalid comments included in the peer feedback received.

We first analyzed students' initial models (before the enactment of peer assessment). The data analysis revealed different levels of increasing sophistication displayed by the students for each component, which align to some of the levels suggested by Papaevripidou et al. (2014). Table 1 shows six levels of increasing sophistication that illuminate the degree of development of the learners' model construction practices, along with the coded student groups assigned to each level. We further analyzed the students' models with respect to the extent to which they drew on the relevant specifications in a valid manner while constructing their models (see table 2).

We then analyzed students' revised models (after the enactment of peer assessment). The revised models of most of the student groups indicated that the students switched to a higher level of attainment in terms of all relevant aspects of their models, including the validity of those aspects (see Tables 1 and 2).

The revised models of most of the student groups indicated that those students switched to a higher level of attainment in terms of all relevant specifications of their models (Representational Power: PP; Interpretational Power: IP; and Predictive Power: PP), including the validity of those aspects. A Wilcoxon rank test showed statistical significant differences between the quality of initial and revised models with respect to the degree to which students had thoroughly addressed the three specifications (RP, IP, and PP) in their models (Z= –3.270; p < .01). Likewise, statistical significant differences were found between the validity of initial and revised models with respect to the RP (Z= –2.0; p < .05) and PP (Z= –3.376; p < .01).

Table 1. Allocation of students' models into different levels of model construction practice following Papaevripidou et al. (2014)

Levels of model construction practice	Level description			Home groups whose models are assigned to each level	
	Representation of the phenomenon	Interpretation of how the phenomenon operates	Predictive power	Initial models	Revised models
Level 1	Superficial[1]	Absent	Absent	11	11
Level 2	Moderate[2]	Absent	Absent	9	-
Level 3	Moderate	Mechanistic[4]	Absent	1, 2, 5	5
Level 4	Moderate	Mechanistic and causal[5]	Absent	-	-
Level 5	Comprehensive[3]	Mechanistic and causal	Limited	3, 4, 6, 7, 8, 10	1, 2, 6, 8, 9
Level 6	Comprehensive	Mechanistic and causal	Strong	-	3, 4, 7, 10

[1] Superficial representation: e.g., most of the components of the phenomenon are missing
[2] Moderate representation: e.g., only few components of the phenomenon are represented
[3] Comprehensive representation: e.g., all components of the phenomenon are represented
[4] Mechanistic interpretation: the model explains how the phenomenon functions
[5] Causal interpretation: the model explains why the phenomenon functions in the way it does

Table 2. Degree of validity for each specification in relation to the levels of model construction practice that emerged from the analysis of students' models

Levels of model construction practice	Validity			Home groups whose models are assigned to each level	
	Representation of the phenomenon	Interpretation of how the phenomenon operates	Predictive power of the model	Initial	Revised
Level 1/2	Invalid	-	-	-	-
	Mostly invalid	-	-	11	11
	Valid	-	-	9	-
Levels 3/4	Mostly valid	Mostly valid	-	5	-
	Valid	Mostly valid	-	-	-
	Valid	Valid	-	1, 2	-
Levels 5/6	Non-valid	Non-valid	Non-valid	-	-
	Mostly valid	Non-valid	Non-valid	-	-
	Mostly valid	Mostly valid	Non-valid	-	-
	Mostly valid	Mostly valid	Mostly valid	3, 4, 7, 8	3, 4, 5, 7, 8
	Valid	Mostly valid	Mostly valid	-	-
	Valid	Valid	Valid	6, 10	1, 2, 6, 9, 10

The type of peer-feedback comments received by assessees was found to be related with the quality of the initial models of the assessees. In particular, negative comments (i.e., references in the peer-feedback comments to what the assessees had not yet achieved) (Kendall's T_b = –

0.373, p <.05) and also justified negative comments (Kendall's T_b = –0.348, p < .05) were related with the quality of assessees' initial models. The data analysis revealed that all revisions (in terms of the student group's attainment of the modeling competence) identified in the revised models of three groups (Groups 3, 8 and 10) were suggested in the peer feedback received, whereas in the revised models of seven groups (Groups 1, 2, 4, 5, 6, 7, and 9) only some revisions were suggested in the peer feedback received. In other words, we identified revisions in the models of seven groups, which were not suggested in the peer feedback comments received from peer assessors. For example, students from Group 1 added an explanation (i.e., *"if green and red coincide, yellow is formed"*) in their revised model which was not included in the peer-feedback comments offered by their peer-assessors (students from Group 2). When examining the initial model of Group 2 (peer-assessors), we detected a sentence resembling the revision of assessee Group 1 (i.e., *"We have 3 sources of light: blue, red and green. If they coincide, one of the colors shown on the figure is formed"*). This is an indication that students from Group 1 might have borrowed this idea while they were assessing the model of Group 2. Triangulation—with data from the post-instructional reflective questionnaire and the interviews with the students—revealed that students proceeded to revising their models, not merely due to the reception of peer feedback comments, but also due to the enactment of the peer assessor role (e.g., engagement to self-reflection processes; exposure to alternative examples while assessing peers' models).

DISCUSSION

This study focused on examining how reciprocal peer assessment in modeling-based learning can serve as a learning tool for learners in a secondary school physics course, in the context of *light and color*. The findings of this study show that reciprocal peer-assessment—experienced by the students in the roles of assessor and assesse—enhanced their learning in the selected topic, as inferred by the quality of their revised models. It is vital to consider that between the model construction and the model revision phase, no instruction or any other kind of intervention took place; therefore, any possible improvements identified in students' revised models arose due to the enactment of peer assessment and in particular either due to the experience that students gained while acting as peer assessors or due to the exploitation of peer feedback received in the peer-assessee role or both.

Students, as assessees, acted on most or all suggestions provided by their peers for revising their models. Students in this study were not reluctant to accept their peers as legitimate assessors, contradicting findings from previous studies (e.g. Tsivitanidou, et al., 2011; Van Gennip, Segers, & Tillema, 2010). They used the peer feedback received from their peer assessors for revising their models and those revisions improved the quality of their models in terms of their RP, IP and PP, as well as in terms of the scientific accuracy of their models. They were able to wisely use the peer feedback received, by filtering peer-feedback comments and finally proceeding with revisions that improved the quality of their initial model with respect to the intended specifications. Students did not proceed with revisions which could potentially undermine the quality of their model. Hence, we have indications of the participants' skills to interpret feedback in a meaningful way and to use to wisely for improving their models. In fact, the analysis revealed that even in cases of receiving invalid feedback comments assessees were

able to filter such invalid comments, as already suggested in previous studies (Hovardas et al., 2014). However, not all revisions detected in assessees' revised models, were explicitly or implicitly suggested in the peer-feedback comments received. We searched for evidence about the possible reasons which might have led assessees to applying those revisions in their models.

The data analysis indicated that assessees revised their models also due to the opportunity which they were offered to act as assessors. In particular, the findings of this study suggest that when students enact the peer-assessor role, they are exposed to alternative examples (i.e., their peers' artefacts) (Tsivitanidou & Constantinou, 2016) which might inspire students to further revise their own artifacts. Also, while enacting the peer-assessor role, the students reconsider the learning objectives that should have been addressed and therefore better appreciate what is required to achieve a particular standard (Brindley & Scoffield, 1998). Moreover, the findings of this study suggest that when students enact the peer-assessor role, they are also engaged in self-reflection processes. Peer assessment, as a process itself, requires self-reflection and in-depth thinking (Cheng, Liang, & Tsai, 2015) and this process bares learning benefits for students. Indeed, students in this study claimed in the post-instructional interviews that while assessing their peers' models, they were engaged in self-reflection processes. They also reported that the opportunity which was given to them to compare—at least implicitly—their own model with that of another peer group while assessing, made them realize what they had interpreted wrongly or not on the basis of their experimental results. This comparison strategy applied by assessors in this study resembles the comparative judgment approach which has been reported as a method that assessors may endorse when offering peer feedback, even if not instructed to do so (Tsivitanidou & Constantinou, 2016).

In this vein, receiving peer feedback while also providing peer feedback was beneficial for students' learning progress. Previous studies in science education have shown that students can benefit from the enactment of peer assessment in terms of their learning (e.g., Prins, Sluijsmans, Kirschner & Strijbos, 2005; Tsai, Lin & Yuan, 2002). The findings of this study suggest that those benefits can also arise in modeling-based learning. We can argue that reciprocal peer assessment can serve as a learning method, confirming findings of previous studies in other contexts and teaching approaches (Orsmond et al., 1996), since students in this study benefited from the reciprocal peer-assessment method, not merely because of receiving peer feedback, but also because they were given the opportunity to act as assessors. The fact that reciprocal peer assessment in modeling-based learning can facilitate students' learning in science, needs to be considered, first, by policy makers and second, by educators, for integrating peer assessment and modeling-based learning in the curriculum and in the everyday teaching practice, respectively.

ACKNOWLEDGEMENTS

This study was conducted in the context of the research project ASSIST-ME, which is funded by the European Union's Seventh Framework Programme for Research and Development (grant agreement no: 321428).

REFERENCES

Acher, A., Arcà, M., & Sanmartí, N. (2007). Modeling as a teaching learning process for understanding materials: A case study in primary education. *Science Education, 91*(3), 398-418.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27,* 427–441.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*, 7–74.

Brindley, C., & Scoffield, S. (1998). Peer assessment in undergraduate programmes. *Teaching in Higher Education, 3*(1), 79–89.

Cestone, C. M., Levine, R. E., & Lane, D. R. (2008). Peer assessment and evaluation in team-based learning. *New Directions for Teaching and Learning, 116,* 69–78.

Chang, H. Y., & Chang, H. C. (2013). Scaffolding students' online critiquing of expert-and peer-generated molecular models of chemical reactions. *International Journal of Science Education, 35*(12), 2028–2056.

Chang, H.-Y., Quintana, C., & Krajcik, J.S. (2010). The impact of designing and evaluating molecular animations on how well middle school students understand the particulate nature of matter. *Science Education, 94*(1), 73–94.

Chen, N.-S., Wie, C.-W., Wu, K.-T., & Uden, L. (2009). Effects of high level prompts and peer assessment on online learners' reflection levels. *Computers and Education, 52,* 283–291.

Cheng, K. H., Liang, J. C., & Tsai, C. C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education, 25,* 78–84.

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research, 83*, 70–120.

Frost, J., & Turner T. (Eds.). (2005). *Learning to teach science in the secondary school: A companion to school experience* (Second Edition), London: Routledge Falmer.

Gielen, M., & De Wever, B. (2015). Scripting the role of assessor and assessee in peer assessment in a wiki environment: Impact on peer feedback quality and product improvement. *Computers & Education, 88,* 370–386.

Gilbert, J. K., & Justi, R. (2016). *Modelling-based teaching in science education.* Springer International Publishing AG Switzerland, ISSN 2213-2260

Grob, R. (2017). *Towards the implementation of formal formative assessment in inquiry-based science education in Switzerland* (PhD Thesis). University of Basel, Switzerland.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81–112.

Hodson, D. (1993). Re-thinking old ways: Towards a more critical approach to practical work in school science. *Studies in Science Education, 22*(1), 85–142.

Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education, 71*, 133-152.

Kim, M. (2005). *The effects of the assessor and assessee's roles on preservice teachers' metacognitive awareness, performance, and attitude in a technology-related design task.* (Unpublished doctoral dissertation). Florida State University, Tallahassee, USA.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.

Labudde, P. (2000): Konstruktivismus im Physikunterricht der Sekundarstufe II (Constructivism in physics instruction at the upper secondary level). Bern, Switzerland: Haupt.

Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug–related tutoring feedback in a computer–based training for written subtraction. *Learning and Instruction, 16*, 310–322.

National Research Council (2007). *Taking science to school: Learning and teaching science in grades K–8*. Washington, DC: National Academy.

National Research Council (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas. Committee on a Conceptual Framework for New K–12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: TheNational Academies Press.

Nelson, M. M., & Schunn, C. D. (2008). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science, 37*, 375–401.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, *31*(2), 199-218.

Nicolaou, C. T., & Constantinou, C. P. (2014). Assessment of the modeling competence: A systematic review and synthesis of empirical research. *Educational Research Review, 13*, 52-73.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of Marking Criteria in the Use of Peer-Assessment. *Assessment and Evaluation in Higher Education. 21*(3), 239–250.

Papaevripidou, M., Nicolaou, C. T., & Constantinou, C. P. (2014). *On Defining and Assessing Learners' Modelling Competence in science Teaching and Learning.* In Annual Meeting of American Educational Research Association (AERA), Philadelphia, Pennsylvania, USA.

Pluta, W.J., Chinn, C.A., & Duncan, R.G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching, 48*(5), 486–511.

Prins, F. J., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J.-W. (2005). Formative peer assessment in a CSCL environment: A case study. *Assessment & Evaluation in Higher Education, 30*, 417–444.

Schwarz, C. V., & Gwekwerere, Y. N. (2006). Using a guided inquiry and modeling instructional framework (EIMA) to support K–8 science teaching. *Science Education, 91*(1), 158–186.

Schwarz, C.V., Reiser, B.J., Davis, E.A., Kenyon, L., Acher, A., Fortus, D., . . . Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, *46*(6), 632–654.

Schwarz, C.V., & White, B.Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. Cognition and Instruction, *23*(2), 165–205.

Sluijsmans, D. M. A. (2002). *Student involvement in assessment, the training of peer assessment skills.* Groningen, The Netherlands: Interuniversity Centre for Educational Research.

Topping, K. J. (2003). Self and peer assessment in school and university reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascaller (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 55–87). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Tsai, C.-C., Lin, S. S. J., & Yuan, S.-M. (2002). Developing science activities through a network peer assessment system. *Computers & Education, 38*(1–3), 241–252.

Tsivitanidou, O., & Constantinou, C. (2016). A study of students' heuristics and strategy patterns in web-based reciprocal peer assessment for science learning. *The Internet and Higher Education. 12,* 12-22. Tsivitanidou, O. E., Constantinou, C. P., Labudde, P., Rönnebeck, S., & Ropohl, M. (2017). Reciprocal peer assessment as a learning tool for secondary school students in modeling-based learning. *European Journal of Psychology of Education*, 1-23.

Tsivitanidou, O. E., & Zacharias, C. Z., & Hovardas, T. (2011). Investigating secondary school students' unmediated peer assessment skills, *Learning and Instruction, 21* (4), 506-519.

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peerassessment as a collaborative learning activity: the role of interpersonalvariables and conceptions. *Learning and Instruction, 20*(4), 280–290.

Van Lehn, K. A., Chi, M. T., Baggett,W., & Murray, R. C. (1995*). Progress report: Towards a theory of learning during tutoring.* Pittsburgh, PA: Learning Research and Development Center, University of Pittsburg.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*, 179–200.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

A TEACHER PERSPECTIVE ON BENEFITS AND CHALLENGES OF PEER-ASSESSMENT

Regula Grob[1,2], Monika Holmeier[1] and Peter Labudde[1]
[1]Center for Science and Technology Education, University of Applied Sciences and Arts Northwestern Switzerland, Basel, Switzerland
[2]University of Teacher Education, Fribourg, Switzerland

Formative assessment has been suggested as a means to support student learning in inquiry-based science education. However, teachers need support in implementing formative assessment practices, such as peer-assessment, in their daily teaching. As a prerequisite for shaping suitable means of support, primary and upper secondary teachers' perspectives on benefits and challenges of peer-assessment in inquiry learning have been explored. Data was collected from 7 primary and 10 upper secondary school teachers from Switzerland who implemented peer-assessment in their science classes. The data included teaching plans, evaluation forms, individual interviews, and group interviews. Inductive coding of the data revealed that the teachers perceived challenges of peer-assessment at the level of teaching practice but also at the level of educational policy. These results suggest that different measures of support such as professional development programmes, but also concrete examples and tools as well as guidelines from educational policy are needed. Considering the benefits of peer-assessment, the teachers from both school levels did not only believe that peer-assessment enhances student learning but also anticipated social and motivational effects. This result implies that formative assessment theories should be more closely connected to learning theories in which student motivation has been identified as a main contributor to learning.

Keywords: formative assessment, peer-assessment, inquiry-based science education

INTRODUCTION

Problem statement

Inquiry and other competence-oriented approaches have become important parts of science education in the recent decades. One issue, however, has been how to support students in their inquiry learning and how to assess respective student competences (e.g. Harlen, 2013). A possible answer to this is the promotion of formative assessment at an international (e.g. OECD, 2005; 2013), but also at a national level (e.g. in the curriculum for the compulsory school levels for the case of Switzerland, D-EDK, 2014). But as a number of studies show, the use of formative assessment in teaching practice varies greatly between teachers (Black, 1993; Bell & Cowie, 2001; Heritage, 2010; Herman, Osmundson & Silver, 2010; Stiggins, Griswold & Wikelund, 1989). The quality of formative assessment rests to a high degree on the strategies teachers use to elicit evidence of student learning and on the use of this evidence to shape subsequent instruction and learning (Bell & Cowie, 2001; Ruiz-Primo, Furtak, Ayala, Yin, & Shavelson, 2010). Subsequently, the need of help for the teachers is stated: "Simply embedding assessments in curriculum does not guarantee improved learning and teaching. Teachers need tremendous support using assessment in their teaching practice" (Yin, et al., 2008, p. 356). The focus of this study will therefore be on science teacher perspectives on peer-assessment, a formative assessment method relatively well-described in the literature (e.g. Topping, 2003), in the context of inquiry learning.

Literature review

Formative assessment has the purpose of assisting learning and for that reason is also called 'assessment for learning'. It involves processes of "seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning and where they need to go and how best to get there" (Assessment Reform Group ARG, 2002, p. 2). The following four characteristic features for an operationalisation of 'formative assessment' were found: (1) Clarity in expectations (e.g. Black, Harrison, Lee, Marshall & Wiliam, 2004); (2) Diagnosis of student level with respect to expectations (Ruiz-Primo et al., 2010); (3) Presence of feedback (Furtak & Ruiz-Primo, 2008) (4) Opportunity to use this feedback (e.g. Andrade, 2010).

For the context of inquiry-based science education, a number of concrete methods of formative assessment have been suggested (e.g. Barron & Darling-Hammond, 2008). The focus of this study will be on peer-assessment which is defined as a process in which students assess their peers' work and provide feedback on it (e.g. Topping, 2003). Peer-assessment follows the idea of "activating students as instructional resources for one another" (Leahy, Lyon, Thompson & Wiliam, 2005, p. 21): Students take both the role of the assessor and the assessee by assessing each other's work. The aim of peer-assessment is to assist peers in identifying the strengths and weakness of their work and to provide suggestions for improving it (Dochy, Segers & Sluijsmans, 1999; Topping, 2003).

A number of advantages and challenges that are associated with peer-assessment have been identified in the literature. The advantages of peer- assessment are, firstly, that feedback from peers who had the same difficulties in the learning progress might suggest direct ways to overcome those difficulties, and formulate them in a language that is naturally used by the students (Black et al., 2004). Secondly, students who assess their peers' work engage in cognitively demanding activities, such as critical thinking (Hanrahan & Isaacs, 2001; Harlen, 2007; Lin, Liu & Yuan, 2001; Lindsay & Clarke, 2001; Topping, 2003; Tsivitanidou, Zacharia & Hovardas, 2011). Thirdly, students get the opportunity to see examples of other students' work. This can potentially lead to self-assessment: By comparing their own work to that of their peers, students can be prompted to reflect on their own learning achievements (Hanrahan & Isaacs, 2001; Lin et al., 2001; Topping, 1998; 2010). Fourthly, peer-assessment may be easier to accept since it is perceived less authoritative than feedback from adults and therefore open to negotiation (Cole, 1991; Topping, 2010). Fifthly, feedback from peers can be more immediate, timely, and individualized than feedback from the teacher (Topping, 2010) simply because there are many more students than teachers in a classroom. Lastly, providing feedback to peers develops the social, communicative, meta-cognitive and other personal and professional skills on the way (Topping, 2010).

Beside the aforementioned advantages, a number of challenges of peer-assessment have also been identified in the literature: When doing peer-assessment, students need to judge the performance of a peer. This needs a certain degree of knowledge in the field that is assessed (Topping, Smith, Swanson & Elliot, 2000). Furthermore, students need to communicate the judgments to their peers and need to provide constructive feedback about their learning process for which communication skills are necessary (Black, Harrison, Lee, Marshall & Wiliam

2003). Thirdly, the recipients need to critically review the feedback and decide on the actions to be taken: Since peer-feedback might include flaws, the recipients need to filter it and then decide whether there is a need to adopt the peers' suggestions and to revise their work (Sluijsmans, 2002). Fourthly, peer-assessment costs lesson time for organization, training and monitoring, particularly in the beginning, if it should be provided at a good level of quality (Topping, 2010). Lastly, social processes influence and contaminate the validity and reliability of assessment provided by peers (Topping, 2010).

Statement of intentions

Following the problem statement, the exploration of the teacher perspective on formative assessment methods such as peer-assessment is considered relevant for a successful implementation of respective approaches. Teachers' perceptions of the benefits and challenges of peer-assessment will therefore be investigated and the implications for supportive measures for the implementation of peer-assessment in inquiry-based science education will be discussed. Furthermore, a widening of the conceptual framework for formative assessment is suggested based on the results.

METHODS

For this study, a 3-semester cooperation with 20 science teachers in Switzerland (9 primary, 11 upper secondary) was established. In every semester, the teachers incorporated a formative assessment method from a pre-defined list (including peer-assessment) in one of their normal inquiry units. The methods were used to assess one or several student competences from another pre-defined list (including, for example, investigation competence, argumentation competence, and modelling competence). The cooperation also included regular meetings with all the teachers, and a teacher manual on the assessment methods which also included illustrative examples.

Data collection

The teachers provided their teaching plans and -materials (student worksheet etc.) from their trials and filled out an evaluation form in which they reflected upon the benefits and challenges of the assessment method. No more than ten days after the trials, individual interviews were held with a sub-group of the teachers (consisting of n=8 teachers from both school levels) in order to speak about the trials and about general issues related to assessment in more detail.

Data analysis

Based on the teaching plans and the teaching materials, it was decided whether the trials included a formative assessment activity. This was evaluated with the four characterizing features of formative assessment as introduced in the literature review. Afterwards, it was decided whether the formative assessment activity was peer-assessment. The respective criterion was whether the students diagnosed and provided feedback on their peers' work. This resulted in 7 primary and 10 upper secondary school cases.

For the analysis of the benefits and the challenges of peer-assessment, the evaluation forms (n=17 evaluation forms) and the transcripts from the individual interviews (n=8 interviews)

were inductively coded. This led to a coding frame with 8 categories for the challenges and 5 categories for the benefits which will be presented in the results part. 18% of the data was double-coded (κ =0.89).

RESULTS

Looking at the challenges, the teachers mentioned difficulties related to the planning of peer-assessment activities (challenge 1). Furthermore, the teachers expressed their doubts about the quality of the diagnosis done by peers (challenge 2), about the quality of the feedback provided by peers (challenge 3), and their uncertainty about their own role (challenge 4). The teachers also anticipated that some of the students might not consider the feedback received from peers to revise their work (challenge 5) or that assessing peers could be boring for students (challenge 6). Another aspect was the role of peer-assessment within the assessment framework, for example the relation between peer-assessment and grading from the teacher (challenge 7). Peer-assessment was also considered rather time-intensive and dependent on a good training of the students (challenge 8).

Considering the benefits, the teachers mentioned that the feedback is provided in a language that is naturally used by the students and it is accepted because the assessor is a peer (benefit 1). Furthermore, the responsibility for the learning in a peer-assessment setting lies with the students, resulting in a lower workload for the teachers and a higher capacity for individual support (benefit 2). The teachers anticipated learning effects in inquiry-specific but also in transversal competences (benefit 3) as well as effects on the classroom climate and the students' motivation (benefit 4). Lastly, the low preparation time for the teacher (benefit 5) was mentioned.

One of the emerging results from the benefits of peer-assessment as mentioned by the teachers is that the teachers from both school levels did not only perceive learning effects (see benefit 3) from peer-assessment but also social and motivational effects (see benefit 4; illustrative quotes: "Peer-assessment enhances the relation between the students"; "Peer-assessment is a way to take students serious and to give value to what they say. This motivates them in their work"). This aspect will be discussed in more detail in the next section of the paper.

DISCUSSION AND CONCLUSIONS

Comparison of the results to the literature

Comparing the benefits and challenges of peer-assessment as mentioned by the teachers in the study to the results found in the literature, a number of aspects are similar. The specific language characteristics of feedback formulated by peers and the responsibility for learning have been previously reported in Black et al. (2004). No references on the resulting capacities of the teachers were found in the research literature, however. The effects of peer-assessment on the students' transversal competences (Topping, 2010) and on self-regulated learning (Hanrahan & Isaacs, 2001; Lin et al., 2001; Topping, 1998; Topping, 2010) have also been previously mentioned but not the effects on the classroom climate and on the students' motivation as anticipated by the teachers in this study. The preparation time was not covered in the literature either.

Considering the challenges, the planning issues as brought up by the teachers in this study are not mentioned in the literature. The quality of the diagnosis (Topping et al., 2000; Topping, 2010) and the quality of the feedback (Black et al., 2003) have been previously discussed. The uncertainty about the own role that resulted, according to the teachers in this study, from the questionable quality of the diagnosis and the feedback, was not found in the literature. The lesson time and the training needed were recognized by Topping (2010), too. None of the teachers in the study spoke about the difficulties in what feedback to use for revision as reported in Sluijsmans (2002).

Overall, the benefits of peer-assessment perceived by the teachers in this study are similar to what is mentioned in the research literature. These effects appear to be independent of the school level and the country-specific context. The social and motivational benefits from peer-assessment have not been found in the literature, though. This will be discussed in more detail in the paragraph 'widening of the theoretical concept needed' below.

The challenges of peer-assessment in the literature were not specifically focussed on the perspective of the teachers nor on organisational issues, resulting in a smaller congruence between the results of this study and the research literature. However, it becomes apparent that the challenges of peer-assessment cannot be neglected.

Support needed

The challenges of peer-assessment appear to need support at different levels to be overcome: Professional development as well as concrete teaching resources could help teachers to enhance their own assessment literacy (see challenges 1, 4) but also to let the students improve their abilities in diagnosing, providing and using peer-feedback (see challenges 2, 3, 5, 6, 8). The role of peer-assessment in the assessment framework (see challenge 7) was the only challenge mentioned that is not situated at the level of teaching and learning practice. Rather, it refers to a more strategic level, with teachers needing help in understanding the relation between formative assessment methods and summative as well as evaluative methods. Guidelines from educational policy representatives could help to clarify the relation between formative and summative assessment.

Widening of theoretical concept needed

Regarding the benefits of peer-assessment, the teachers did not only perceive learning effects but also social and motivational effects. This is not aligned with formative assessment theory which focusses on the former by conveying the idea that formative assessment supports student learning (Black & Wiliam, 1998; Natriello, 1987). Interdependencies between formative assessment and student motivation (Black & Wiliam, 1998) and a relation between formative assessment and student confidence (Smit, 2009) have been suggested, but literature on these effects is generally scarce. The result suggests that the formative assessment theory should be widened towards learning theories in which student motivation has been identified as a main contributor to student learning.

Retrospects and prospects

The aim of this study was to explore teachers' perceptions on benefits and challenges of peer-assessment in order to shape suitable means of support for teachers. The study was conducted with a small number of participants and in an open setting where the teachers designed the inquiry units themselves. It is therefore hard to decide on the specifity of the results (e.g. to what extent the challenges refer to peer-assessment specifically rather than to formative assessment methods in general). Nevertheless, the participating group of teachers included different school levels, subjects, years of teaching experience and gender. Furthermore, the rich data on the teachers' trials and their reflections upon them provide a dense picture of the teachers' perspectives on peer-assessment in the context of inquiry.

The study results in two main outcomes: Firstly, it offers first ideas on how to support the uptake of more peer-assessment in daily teaching practice. Secondly, it provides implications on how to further develop formative assessment theories.

ACKNOWLEDGEMENT

The work presented in this paper is part of the ASSIST-ME project which is funded by the European Commission (Seventh Framework Programme for Research; grant agreement no: 321428).

REFERENCES

Andrade, H. (2010). Students as the definitive source of formative assessment. In H. Andrade & G.J. Cizek (Ed.), *Handbook of formative assessment.* New York: Routledge.

ARG (Assessment Reform Group) (2002). *Assessment for learning: 10 Principles.* London: ARG.

Barron, B. & Darling-Hammond, L. (2008). Teaching for meaningful learning: A review of research on inquiry-based and cooperative learning. In L. Darling-Hammond, B. Barron, P. D. Pearson, A. H. Schoenfeld, E. K. Stage, T. D. Zimmermann, G. N. Cervetti, & J. Tilson (Eds.), *Powerful Learning. What we know about teaching for understanding* (pp. 11-70). San Francisco: Jossey-Bass.

Bell, B. & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*(5), 536–553.

Black, P. (1993). Formative and summative assessments by teachers. *Studies in Science Education, 21,* 49-97.

Black, P., Harrison, Ch., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* London: Open University Press.

Black, P., Harrison, Ch., Lee, C., Marshall, B., & Wiliam, D. (2004). *Working inside the black box: Assessment for learning in the classroom.* Phi Delta Kappan.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice. 5*(1), 7-73.

Cole, D.A. (1991). Change in self-perceived competence as a function of peer and teacher evaluation. *Developmental Psychology, 27,* 682-688.

Deutschschweizer Erziehungsdirektoren-Konferenz D-EDK (2014). *Lehrplan 21* [Curriculum 21]. Luzern: D-EDK.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-peer and co-assessment in higher education: A review. *Studies in Higher Education, 24(3),* 331-350.

Furtak, E. M., & Ruiz-Primo, M. A. (2008). Making students' thinking explicit in writing and discussion. An analysis of formative assessment prompts. *Science Education, 92*(5), 799-824.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research and Development, 20,* 53–70.

Harlen, W. (2007). Holding up a mirror to classroom practice. *Primary Science Review, 100,* 29–31.

Harlen, W. (2013). *Assessment & inquiry-based science education: Issues in policy and practice.* Trieste: Global Network of Science Academies (IAP) Science Education Programme (SEP).

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom.* Thousand Oaks, California: Corwin Press.

Herman, J. L., Osmundson, E., & Silver, D. (2010). *Capturing quality in formative assessment practice: Measurement challenges, CRESST Report 770.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Leahy, S., Lyon, Ch., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Assessment to promote learning, 63(3),* 19-24.

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: Feedback for students with various thinking styles. *Journal of Computer Assisted Learning, 17,* 420-432.

Lindsay, C., & Clarke, S. (2001). Enhancing primary science through self– and paired–assessment. *Primary Science Review, 68,* 15–18.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22,* 155-175.

Organization for Economic Co-operation and Development OECD (2005). *Formative assessment. Improving learning in secondary classrooms.* Paris, France: OECD Publishing.

Organization for Economic Co-operation and Development OECD (2013). *Synergies for better learning: An international perspective on evaluation and assessment.* OECD Reviews of Evaluation and Assessment in Education. OECD Publishing, Paris.

Ruiz-Primo, M. A., Furtak, E. M., Ayala, C., Yin, Y, & Shavelson, R. J. (2010). Formative assessment, motivation, and science learning. In H. Andrade & G. J. Cizek (Ed.), *Handbook of formative assessment* (pp. 139 – 158). New York: Routhledge.

Sluijsmans, D. M. A. (2002). *Student involvement in assessment, the training of peer-assessment skills.* Maastricht: Interuniversity Centre for Educational Research.

Smit, R. (2009). *Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz. Eine empirische Studie in der Sekundarstufe 1 [Formative assessment and its use for the development of learning competence. An empirical study at lower secondary school level].* Schneider Verlag Hohengehren GmbH: Baltmannsweiler.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68,* 249–276.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Cascaller (Eds.), *Optimising new modes of assessment: in search of qualities and standards* (pp. 55–87). The Netherlands: Kluwer Academic Publishers.

Topping, K. J. (2010). Peers as a source of formative assessment. In H. Andrade & G.J. Cizek (Eds.), *Handbook of formative assessment* (pp. 61- 74). New York: Routledge.

Topping, K., Smith, F. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment and Evaluation in Higher Education, 25,* 149-169.

Tsivitanidou, O., Zacharia, Z. C., & Hovardas, A. (2011). High school students' unmediated potential to assess peers: Unstructured and reciprocal peer assessment of web-portfolios in a science course. *Learning and Instruction, 21,* 506-519.

Yin Y., Shavelson, R.J., Ayala, C.C., Araceli Ruiz-Primo, M., Brandon, P. R., Furtak, E. M., Tomita, M. K., & Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335-359.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

DEVELOPMENT AND VALIDATION OF LEARNING PROGRESSIONS ON CHEMICAL CONCEPTS

Kübra Nur Celik and Maik Walpuski
University of Duisburg-Essen, Essen, Germany

The development of scientific literacy is very important for lifelong learning and the understanding of core concepts in science (AAAS, 2007). At the same time, a study conducted in North Rhine-Westfalia in Germany as a national assessment (Pant et al., 2013) shows that a lot of students perform only poorly in standardized assessment tests in chemistry and do not even reach the necessary basic skills. These students often lose track in chemistry instruction because of their early knowledge deficits and inability to catch up accordingly. To support these low-achievers it is important to investigate how essential ideas and concepts are related to each other and how they contribute to the logical (in large parts hierarchical) structure of chemical knowledge. For the German context learning progressions for the chemical concepts "Structure of Matter", "Chemical Reaction" and "Energy" (c.f. MSW, 2011) for the first two learning years in chemistry instruction have been developed, with several core ideas and their specific requirements. The first aim of the presented project is to evaluate these learning progressions empirically. In addition, it focuses on defining achievable minimal knowledge levels that guide all students to gain scientific proficiency in the long run. On the basis of performance tests specific to the assumed learning progressions it is possible to identify interdependencies between the core ideas and evaluate the progressions' validity. The pilot study reported here primarily describes the test instrument, its test parameters and possible methodological considerations for analyzing the main study data, which is not yet complete.

Keywords: learning progressions, competencies, chemical concepts

INTRODUCTION AND THEORETICAL FRAMEWORK

Problem and initial situation

Similar to other nations, Germany has introduced educational standards, which describe competencies the students should have acquired by the end of a particular grade (KMK, 2005). These educational standards are formulated as general standards addressing the average performance level (Klieme et al., 2007). However, the 2012 IQB national assessment study (Pant et al., 2013) revealed that German students, particularly in North Rhine-Westphalia, perform lowly on these standardized assessment tests in chemistry. With regard to an US study (Alonzo & Gotwals, 2012) it can be assumed that low test results may be related to unfocused and disconnected science education. The reason for this could lie in the largely hierarchical structure of chemistry knowledge. The hierarchical structure might put students, particularly low-achieving ones, who lost track at some point during chemistry instruction, at a disadvantage, where they are unable to catch up on the content. In order to support these students it is necessary to investigate the relationship between essential ideas and concepts in chemistry and their contribution to meaningful learning and knowledge structures. One possible approach is to map the interdependencies as learning progressions and use them as a guiding framework for structuring chemistry instruction within the first two learning years in chemistry. Teachers might also use the learning progressions to identify difficulties

understanding concepts and ideas on an individual basis and derive according supporting measures.

Theoretical framework

This study uses the concept of learning progressions as a way of describing the structure of chemical content knowledge. Learning progressions propose the development of essential core ideas that support cross-linked knowledge and can be read as possible learning pathways to develop professional competencies. They also postulate a particular sequence of abilities and core concepts, which students have to acquire over time (e.g. Corcoran, Mosher, & Rogat, 2009; Duschl, Schweingruber, & Shouse, 2007; Duncan & Hmelo-Silver, 2009; Stevens, Delgado, & Krajcik, 2009).

Learning progressions consist of several core ideas the students have to understand. Students enter the progression with their prior knowledge and abilities (lower anchor). They proceed through predetermined learning pathways successively to achieve the learning targets which describe skills and knowledge for end of the progression (upper anchor) (Corcoran, Mosher, & Rogat, 2009). The levels between the lower and upper anchor are defined by the learning performances which set the level of understanding and competencies students would be able to perform (Corcoran, Mosher, & Rogat, 2009; Duncan & Hmelo-Silver, 2009).

Other studies have already used learning progressions successfully. The American Association for the Advancement of Science (AAAS, 2007), for instance, aspires in “Project 2061” the idea of developing scientific literacy for all students and developed learning progressions for various domains in science education, such as Physical Science and Earth Science. They used strand maps to visualize the development of students’ understanding of core ideas at different stages of progress and represent the link between core ideas and learning targets to diagnose students’ conceptual abilities (AAAS, 2007). There have also been first attempts at developing and validating a learning progression via strand maps for the concept of energy in physics in the German context (Neumann, Viering, Boone, & Fischer, 2013). In addition, first investigations of core ideas related to the basic concept “Structure of Matter” and “Chemical Reaction” in chemistry have already been conducted, as well (Weber, Emden, & Sumfleth, 2016). However, rare attention has been paid to a learning progression for all three basic concepts in chemistry and the interdependencies of their core ideas.

PROCEDURE AND DESIGN

Research questions

The following research questions are addressed by this study:

1. Can the developed Learning Progression be validated empirically?
2. Is there an interdependency between the chemical concepts? Are requirements from one chemical concept necessary to achieve requirements from a different chemical concept?

Study context and preliminary work

In a quasi-longitudinal study, students in the first two learning years in chemistry instruction at comprehensive schools in North Rhine-Westphalia are tested. Prior to testing, a working group consisting of science education researchers, school teachers and educational administration stakeholders has developed a preliminary strand map and its core ideas as anchors. On the basis of educational standards for chemical education (KMK, 2005), school books and school curricula this team has identified 57 core ideas for the three chemical concepts "Structure of Matter", "Chemical Reaction" and "Energy" (c.f. MSW, 2011) for the first two learning years in chemistry (Table 1). This is the equivalent of grades 8 and 9 at the lower secondary level in Germany.

Table 1. Distribution of the developed 57 core ideas across the three chemical basic concepts for the first two learning years in chemistry.

	1st learning year	2nd learning year
Structure of Matter	13	19
Chemical Reaction	6	7
Energy	7	5

Each core idea is framed by a description of what students are expected to know and be able to do if they have fully understood the core idea. Additionally, boundaries were formulated describing what students are not expected to know at this point. Usually these boundaries are defined by content of another core idea or the complexity of the content idea for this level). Typical misconceptions of students are also related to the core ideas and can be used as distractors in the assessment test (Figure 1).

These chemical core ideas were then brought into a logical sequence and were connected via stand maps (analogous to the project of AAAS (2007)) (Figure 2).

The strand map considers the hierarchical arrangement of the core ideas over the first two learning years and differentiates between necessary and sufficient requirements for a meaningful construction of knowledge. Requirements, which are assumed to be necessary for the understanding of the hierarchically higher core idea are represented with red arrows and the requirements, which are not assumed to be necessarily relevant for the hierarchically higher core idea are represented with black arrows (Figure 2).

Basic concept: Structure of Matter	**Learning year:** ☐1 ☒2

Core idea: Protons and neutrons can be found in the atomic nucleus and constitute almost the whole mass of an atom, while electrons are located in the electron shell and determine the size of an atom (Rutherford).

Expectations: Students are expected to know that ...

- forces act between the elementary particles of an atom.
- the electrons build the atomic shell.
- the protons and neutrons build the atomic nucleus.
- the mass of an atom is almost completely determined by the atomic nucleus.
- the size of an atom is determined by the atomic shell.
- an atom predominantly is void.
- proton and neutron each have a mass of one u.

Boundaries: Students do not have to know (for this core idea) ...

- that the mass of proton and neutron are marginally different.
- which influence the electron has on the mass of an atom.

Typical misconceptions:

- The atomic shell contains air.
- The atomic shell is an actual shell.

Figure 1. Description of a core idea.

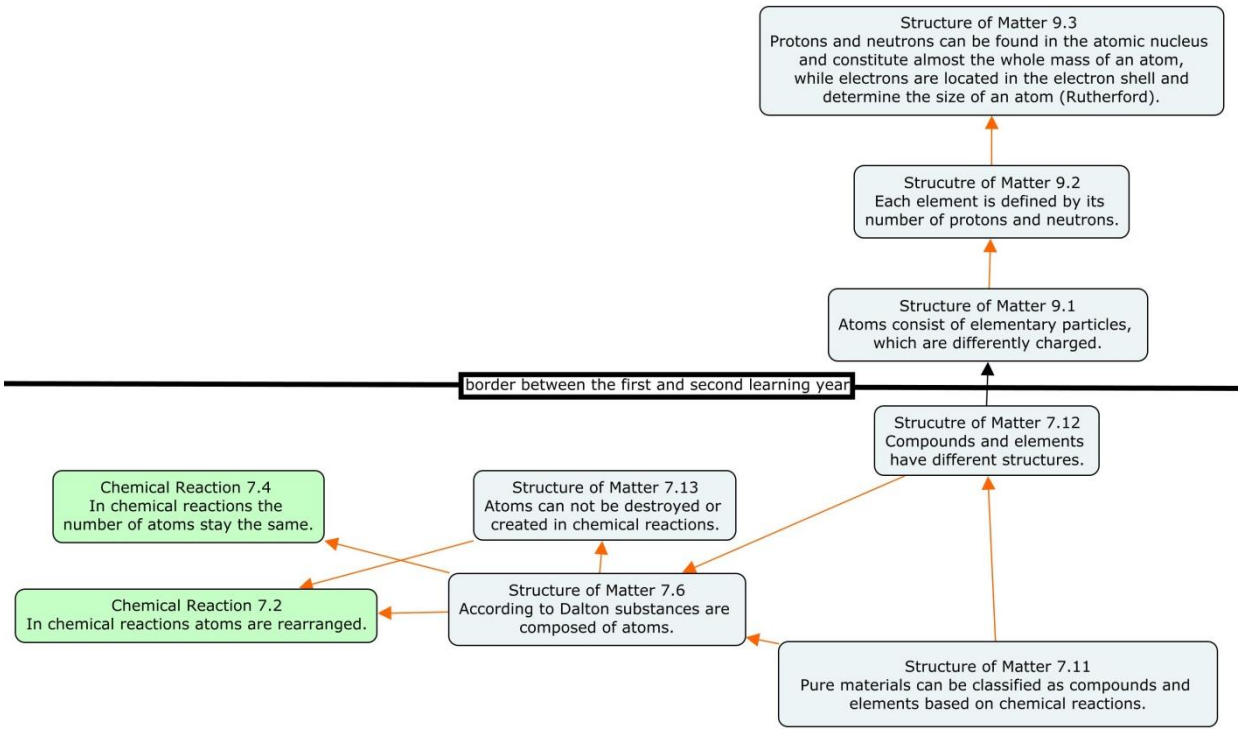


Figure 2. Exemplary part of the strand map for the chemical basic concepts "Structure of Matter" and "Chemical Reaction" for the first two learning years.

Test items and test design

In order to test the validity of this purposed strand map as well as to assess students' abilities it was necessary to develop suitable test items. The test items reflect each core idea and its according expectations which the students are expected to know. For a sufficiently large item pool at least five test items have been developed for each of the core ideas (Table 1).This resulted in a total of 329 items in a multiple-choice single-select format. In the pilot study this item pool was used to identify problematic items which should be removed for a final test instrument. Due to the large number of items, a multi-matrix design was realized, where the test items were distributed among 25 different test booklets. The test booklets were constructed by considering the relations between the core ideas in the strand map (Table 2).

Table 2. Approach to construct a test booklet for the exemplary part of the strand map in Figure 2

Test booklet	Items for the main core idea	Items for the directly connected core ideas
1	Structure of Matter 7.6	Structure of Matter 7.11, 7.12, 7.13; Chemical Reaction 7.2, 7.4

Note: The core ideas Structure of Matter 7.13 and Chemical Reaction 7.2 are also represented as main core ideas because all items of the directly connected core ideas are in this test booklet.

In consequence, not all the 25 test booklets contain the same amount of items because of the differing number of relations in the strand map between core ideas. While constructing the test booklets it was also necessary to pay attention to items giving the answer to succeeding ones. In order to achieve a true multi-matrix design, each test booklet was anchored via overlapping items of at least one core idea so that an overall analysis of all test items was ensured. The test items were administered to 787 students from grades 8 to 10. In the German school system these grades typically correspond to the first three learning years in chemistry. Grade 10 (third learning year) students were tested additionally by intention to obtain data from students who are expected to know all core ideas and to generate sufficient variance in the performance of the students.

Methods

The basis for the following analyses are unidimensional Rasch models from item response theory as it is expected that items are not equally difficult to solve. To make valid statements about the quality and reliability of the test items, they were analyzed with regard to their test parameters and model fit parameters. These items were also ordered with increasing item difficulty on a Wright Map to get a first rough estimation of whether the whole item difficulty spectrum is covered for all three basic concepts and items for hierarchically higher core ideas (second learning year) are more difficult than lower ones (first learning year). All analyses were conducted using ConQuest® software (Wu, Adams, & Wilson, 2007).

Results of the pilot study

The following results refer to the pilot study data of N = 787 students (50.3 % female). Of this sample, 33.7 % are from the first learning year, 53.1 % from the second learning year and 13.2 % from the third learning year. The students should work on items whose item solution they know. So that all of the crude and wrong answered items were assessed as false. Besides, the items would not be answered in the same number so that in an incomplete block design for the core ideas each item reached 32 responses in average. The following table (Table 3) presents the model fit parameters of all 329 multiple-choice single-select items. All, but are within the weighted-Mean-Square threshold value between 0.80 and 1.20. In addition, the t-statistics for all but three fall within the tolerable range of $-1.98 < t < 1.98$. The item reliability is excellent so that the item difficulties are estimated accurately. The EAP/PV reliability is also satisfying (0.828), which means the estimated person abilities are accurate, as well.

Table 3. Fit statistics of the 329 test items.

Items	EAP/PV Reliablity	Item Reliablity	Item Difficulty	wMNSQ	*t*-statistics
329	0.828	0.913	-1.936 - 3.954	0.8 - 1.22	-3.4 – 3.3

Items with problematic fit measures were analyzed in more detail via distractor analyses. An observation of the according item characteristic curves and their item discrimination values revealed that they had anomalous curve patterns and therefore should be revised or removed from the test instrument.

The item difficulty varies between M = 0.6169 (SD = 0.0779) logits for the first and M = 1.3096 (SD = 0.0613) logits for the second learning year. A paired t-test reveals a significant difference between them with a medium-sized effect ($t(297.841) = -6.981$, $p \leq .001$, $d = 0.77$). Hence, items for the second learning year are significantly more difficult than items for the first year.

In the strand map the core ideas are hierarchically arranged. Therefore, the item difficulties are expected to be different for the first two learning years. As can be seen in the Wright Map (Figure 3) the item difficulties and person abilities are normally distributed. However, the difficulty of the items is above average for the students. The three basic concepts consist of difficult items as well as easy items, but easy items for low-ability persons are missing. It is assumed that the mismatch between person ability and item difficulty is due to fact that some of the content has not been covered by the teacher or that the low-achieving students are left behind at some point and are not able to follow anymore.

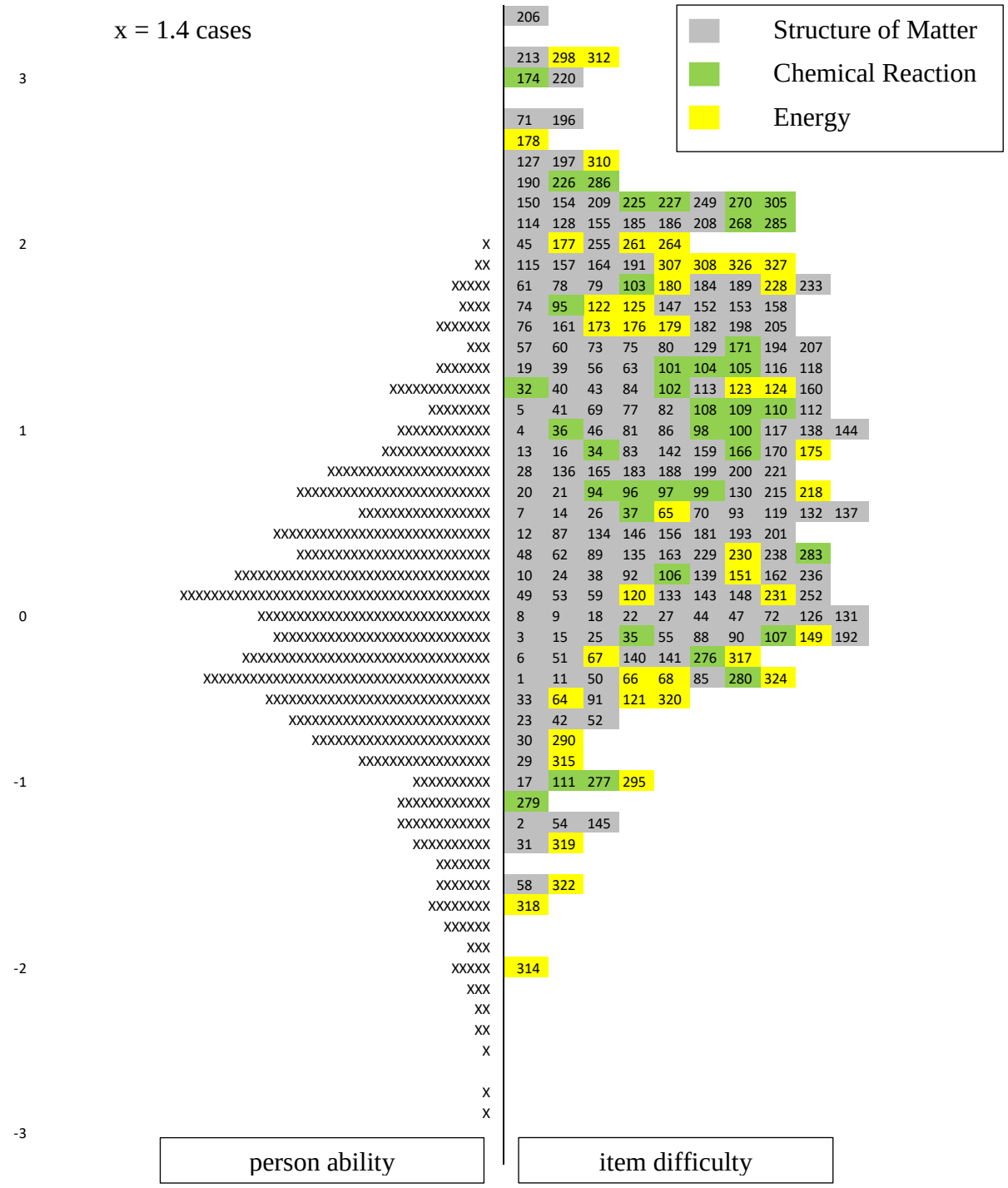


Figure 3. Wright Map for all test items of the pilot study.

The mean value for the items of the core ideas from Chemical Reaction is M = 1.0890 (SD = 0.1033), for the items of Energy M = 0.9732 (SD = 0.1485) and for the items of Structure of Matter M = 0.9619 (SD = 0.0638) (Figure 4).

An ANOVA revealed that the item difficulties of items for the three basic concepts are not significantly different from one another ($F(2, 326) = 0.454$, $p = .636$, $\eta^2 = .003$). All of these analyses show that the test items are suitable for our investigations and can be used in the main study.

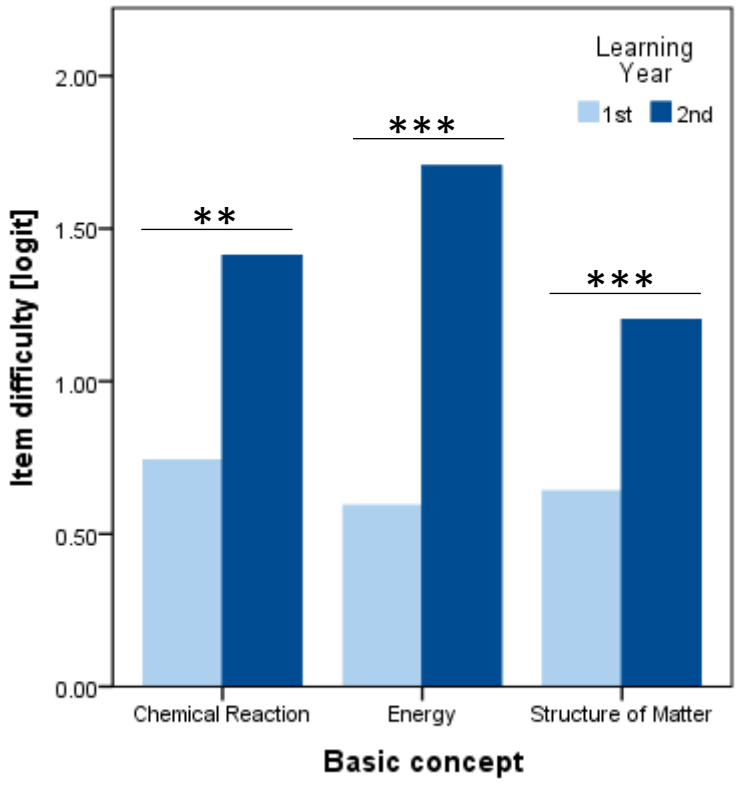


Figure 4. Comparison of the items difficulties for the three basic concepts.

OUTLOOK FOR THE MAIN STUDY

In the main study the revised performance test will be administered at two points of measurement (in the middle and at the end of a school year) to investigate the hypothetical interdependencies between the core ideas. For each interdependency between two core ideas, approximately 50 answers per item and point of measurement are needed.

The following example shall illustrate the methodological approach to verify or falsify the interdependencies between core ideas (Figure 5): Core idea A is (hypothetically) the requirement for understanding core idea B. The items for both core ideas (A and B) will be administered to the same students. As a consequence, the dependency between the core ideas A and B can be tested by analyzing solution probabilities. Ideally, all students who answer the items for the core idea B correctly also answered the items for idea A correctly, so that the dependency between the two core ideas is verified. In the other extreme case all B-solvers did not answer the items for the core idea A correctly, in which case the dependency is disproven. Certainly mixed cases are also possible, which have to be determined by a quantitative threshold.

There is no standardized procedure to test learning progressions. Therefore statistical methods with a different focus (time, person, items) like the cross-lagged panel analysis, the McNemar test, the Guttman scale, and the Bayesian network will be used as possible methodical ways to test the hypothetical assumptions made in the strand map.

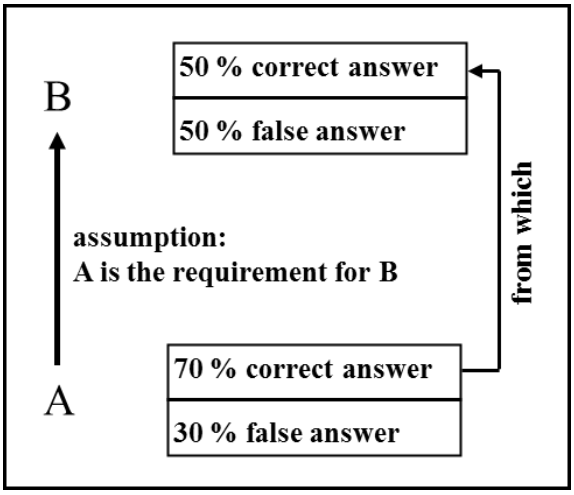


Figure 5. Example for the methodological approach.

The cross-lagged panel analysis enables to predict the performance during later points of measurement on the basis of the first performance data at the first points of measurement (Kenny, 1975; Döring & Bortz, 2016). The McNemar-test examines whether the item for core idea B are more difficult to solve than items for core idea A at several measurement points and allows to divide the students into the groups "solved the item" and "did not solve the item" (Eid, Gollwitzer, & Schmitt, 2013; Field, 2014).The Guttman scale ranks test items as indicated by their solution probability and shows which students are able to solve the items in the basis of their ability. The students who solve the more difficult items also solve the easier items for the same content (Döring & Bortz, 2016). The Bayesian networks investigate the overall hierarchical structure of several interdependencies between the connected core ideas in the strand map because it focuses on conditional probabilities to evaluate if one core idea is conditional on the probability of the other core ideas (Mislevy & Gitomer, 1996 in West et al., 2012).

The results of the main study can be used as evidence about the necessity of one chemical core idea to understand the next one or whether knowledge in one idea is just beneficial for the understanding of the others. Therefore, the results of the study should enable to diagnose students' deficits so that teachers can explicitly support particularly low-achieving students to reprocess their deficits by working off the relevant chemical core ideas, which are based on each other and are indispensable for the construction of systematic knowledge. Learning progressions promise to build a better connecting point between standards, curriculum, instruction and assessment to improve science education and to promote scientific literacy (Alonzo & Gotwals, 2012; Duncan & Hmelo-Silver, 2009). So instruction can be better coordinated and student learning can be supported target-oriented.

ACKNOWLEDGEMENT

Many thanks to the SINUS working group and QUA-LiS who support the project and also the students who participated in the performance tests.

REFERENCES

Alonzo, A. C., & Gotwals, A. W. (2012). Introduction: Leaping into Learning Progressions in Science. In A. C. Alonzo & A. W. Gotwals (Eds.). Learning Progression in Science, Current Challenges and Future Directions (pp. 3-12). Rotterdam: Sense Publishers.

American Association for the Advancement of Science (AAAS) (2007). Atlas of Science Literacy. Volume 2. Washington, DC: AAAS.

Corcoran, T., Mosher, F. A., & Rogat, A. (Eds.) (2009). Learning Progressions in Science. An Evidence-based Approach to Reform. Philadelphia, PA: CPRE.

Döring, N., & Bortz, J. (2016). Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. [Research methods and evaluation in social and human sciences]. Heidelberg: Springer.

Duncan, R. G., & Hmelo-Silver, C. (2009). Editorial – Learning Progressions: Aligning Curriculum, Instruction, and Assessment. *Journal of Research in Science Teaching*, *46*(6), 606-609.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.) (2007). Taking Science to School: Learning and Teaching Science in Grades K-8. Washington, DC: The National Academies Press.

Eid, M., Gollwitzer, M., & Schmitt, M. (2013). Forschungsmethoden. [Research methods]. Weinheim: Beltz.

Field, A. (2014). Discovering Statistics using IBM SPSS Statistics. Los Angeles, London, New Delhi, Singapore, Washington DC: Sage Publications.

Kenny, D. A. (1975). Cross-Lagged Panel Correlation: A Test for Spuriousness. *Psychological Bulletin*, *82*(6), 887-903.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riwuarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. (2007). Zur Entwicklung nationaler Bildungsstandards – Expertise (For the development of national educational standards – Expertise). Bonn, Berlin: BMBF.

KMK, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Eds.) (2005c). Bildungsstandards für das Fach Chemie für den Mittleren Schulabschluss. [Educational standards for the chemical standards for middle school]. München: Luchterhand.

Ministerium Für Schule und Weiterbildung NRW (MSW) (2011). Kernlehrplan für die Gesamtschule – Sekundarstufe I in Nordrhein-Westfalen. Naturwissenschaften. Biologie, Chemie, Physik. [Core curriculum for the comprehensive school - Middle school in North Rhine-Westphalia. Sciences. Biology, Chemistry, Physics]. Frechen: Ritterbach Verlag.

Neumann, K., Viering, T., Boone, W., & Fischer, H. E. (2013). Towards a Learning Progression of Energy. *Journal of Research in Science Teaching*, *50*(2), 162-188.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.) (2013). The IQB National Assessment Study 2012. Competencies in Mathematics and the Sciences at the End of Secondary Level I. Münster, New York, München, Berlin: Waxmann.

Stevens, S. Y., Delgado, C., & Krajcik, J. S. (2009). Developing a Hypothetical Multi-Dimensional Learning Progression for the Nature of Matter. *Journal of Research in Science Teaching*, *47*(6), 687-715.

Weber, K., Emden, M., & Sumfleth, E. (2016). Development of a Learning Progression on Chemical Reactions. In J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto & K. Hahl (Eds.), Electronic Proceedings of the ESERA 2015 Conference. Science education research: Engaging learners for a sustainable future, Part Evaluation and assessment of student learning and development/11 (co-ed. J. Dolin & P. Kind), (pp. 1589-1597). Helsinki, Finland: University of Helsinki.

West, P., Wise Rutstein, D., Mislevy, R. J., Liu, J., Levy, R., Dicerbo, K. E., Crawford, A., Choi, Y., Chapple, K., & Behrend, J. T. (2012). A Bayesian Network Approach to modelling Learning Progressions. In A. C. Alonzo & A. W. Gotwals (Eds.). Learning Progressions in Science, Current Challenges and Future Directions (pp. 257-292). Rotterdam: Sense Publishers.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). ACER ConQuest version 2.0: generalised item response theory modelling software. Camberwell, Vic: ACER Press.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

INCLUSION IN CHEMISTRY EDUCATION IN SECONDARY SCHOOL

Dagmar Michna and Insa Melle
TU Dortmund University, Chair of chemical education, Dortmund, Germany

The UN-Convention on the Rights of Persons with Disabilities (CRPD) from 2009 requires the right of equal participation in schools for students with and without special educational needs (CRPD, 2006). In accordance with this convention, the school law act of North Rhine-Westphalia was amended in 2013 (NRW, 2013). The demand for inclusion does not mean that the curriculum has to be designed entirely unique, but that the students work on the same content individually (Kullmann, Lütje-Klose & Textor, 2014). The implementation of inclusive teaching is difficult, as there are very few and insufficient learning environments, especially in the field of science. In order to find a more efficient method of implementing inclusive teaching, we developed a concept that combines instructive as well as constructive elements, and the Universal Design for Learning (UDL, CAST, 2011). The main idea is to involve all students in the learning process by offering varied ways to access a certain content. The aim of this study is to develop and evaluate an inclusive teaching unit in chemistry (Michna, Melle & Wember, 2016). On the one hand, it contains a lecture given by the teacher and on the other hand, the learners use a self-evaluation sheet in order to identify their own learning abilities and their aspired proficiency levels. Learners first asses themselves on a four-point Likert scale to illustrate what they have already learned from the lecture. Afterwards, the students decide what knowledge they want to achieve. Then, they work with material that is based on the UDL. The study is carried out with two different groups of secondary education students (Grade 8, n = 172). Both groups deal with the same material in a 225-minute inclusive teaching unit. The difference between the groups is their composition: The intervention group is an inclusive learning group, while there are no students with disabilities in the control group.

Keywords: inclusion, universal design for learning, chemistry education

MOTIVATION

In June 1994 representatives of 92 governments and 25 international organizations formed the World Conference on Special Needs Education, held in Salamanca, Spain (United Nations Educational, Scientific and Cultural Organization, 1994). The conference established a new framework, approving that ordinary schools should accommodate all students regardless of their physical, intellectual, or social background. Germany ratified the Convention on the Rights of Persons with Disabilities in 2009. But even today, there are still few and insufficient learning environments, especially in the field of science. As a consequence, the goal of the study presented in this paper is to develop and evaluate an inclusive learning unit in chemistry for secondary schools. Therefore, this project combines instructive elements, constructive learning phases, and the UDL.

THEORETICAL BACKGROUND

Instruction and construction

An example of instruction is the *direct instruction* (Engelmann, 1980), which is based on the assumption that every student can do well, if he receives proper instructions. Direct instruction

implies a teaching concept that serves to learn basic knowledge in which teaching is intended teacher-centered (Grell, 1999; Gruehn, 2000, pp. 42; Hasselhorn & Gold, 2009, pp. 241; Quittenbaum, 2016). Furthermore, an example of construction is the *self-regulated learning* (Zimmerman & Martinez-Pons, 1988). This method includes the use of self-regulated learning strategies, students' responsiveness for self-oriented feedback, and the motivation to achieve academic goals which are personally intended by the students. Evidence shows that self-regulated learning can lead to greater success in learning (Pintrich & De Groot, 1990) and that accurate instruction has positive effects on learning outcomes (Touvinen & Sweller, 1999; Klahr & Nigam, 2004).

Universal Design for Learning

As a result of the heterogeneity in classes, the design of learning environments has to be changed, as different aspects have to be taken into account regarding planning, implementation and analysis of lessons. The Universal Design for Learning (UDL) is a framework for the design of inclusive learning environments that has been proposed in the US as being an evidence-based approach to make schools and learning accessible for all learners. The leading idea is that successful learning for all may only be possible if all students have access to the learning content. (Center for Applied Special Technology [CAST], 2011; Meyer, Rose, & Gordon, 2014). In detail, the framework of UDL consists of instructional approaches that provide students choices and alternatives concerning the materials, contexts, contents etc. A successful learning environment supports and challenges students while minimizing barriers. Minimizing barriers requires flexible teaching methods and materials. Accordingly, the UDL framework consists of three overarching principles (CAST, 2011):

1. "Principle I: Provide Multiple Means of Representation (the "what" of learning)
2. Principle II: Provide Multiple Means of Action and Expression (the "how" of learning)
3. Principle III: Provide Multiple Means of Engagement (the "why" of learning)"

To go more into detail, the principles are broken down into guidelines and checkpoints. The UDL can be summarised in a table, where the guidelines, the three principles and the check-points are given (Table 1). Guideline 2, for example, deals with options for language or symbols. A picture or image that carries a specific meaning for some learners may carry a very different meaning for other learners from different cultural backgrounds. As a result, inequalities can arise when information is presented through a single form of representation.

By implementing the UDL it should be possible to reduce barriers in methods and materials, and to provide access to information and learning, ideally for all students.

RESEARCH QUESTIONS

Based on the theoretical background, the question arises whether a learning environment, which consists of instructive and constructive elements and which is designed by using the UDL, leads to a comparable knowledge growth of all learners in inclusive and non-inclusive classrooms.

Thus, this study addressed the following research questions:

1. Is the increase in knowledge of both groups comparable?
2. Is the increase in knowledge of both groups comparable in the long term?
3. Is the teaching unit rated as equally well by both groups?

Table 1. Universal Design for Learning (CAST, 2011)

I Provide Multiple Means of Representation	II Provide Multiple Means of Action and Expression	III Provide Multiple Means of Engagement
Perception	Physical action	Recruiting interest
Language, expressions, and symbols	Expression and communication	Sustaining effort and persistence
Comprehension	Executive functions	Self-regulation

DESIGN

The procedure of the main study is based on the results of a pilot study. The following illustrations only refer to the main study.

The learning unit deals with the topic “chemical reaction” which consists of five 45-minutes lessons and is a new topic for all of the students. To answer the research questions, two experimental groups were created. Both groups work with the same materials during the whole time. The major difference between the groups is their composition: The intervention group 1 (WithinSEN) is an inclusive learning group, while there are no students with special needs in the intervention group 2 (WithoutSEN). One week before the learning unit, chemistry performance, intellectual performance and academic self-concept are assessed. Furthermore, the student skill assessment is compiled by using a rating by the teacher (pre-test, 60 minutes). The first lesson starts with a 10-minute lecture given by the teacher. After the lecture, the students work with the self-evaluation sheets and the learning materials. These two lessons are followed by an experiment-based lesson. At the beginning of the experimental phase, a short safety briefing is conducted, as most students of the participating classes have no experience in experimenting. Finally, the last two lessons contain the combination of a lecture given by the teacher and also of self-regulated work again. One week after the learning unit, the chemistry performance is measured again and the additionally, students’ feedback is assessed (post-test, 45 minutes). Four weeks after the second measure-point the chemistry performance is collected for the third time (follow-up-test, 30 minutes).

METHODS AND MATERIAL

Lecture

The lectures are supported by power-point-presentations and provide first information about the topic. Both lectures have a timeframe of approximately10 minutes and are given by the teacher at the beginning of the self-regulated workphases. Both lectures include three subtopics on the topic of chemical reaction. In total, the first power-point-based lecture consists of 24 slides, which also contain explanations on the work with the self-evaluation sheets. In comparison, the second lecture contains only 19 slides.

Each subtopic is discussed in a similar way within each lecture as each of them consists a start-up slide illustrating the focused question, followed by the explanation of the content and ending with a summary.

Self-evaluation-sheet

The self-evaluation-sheets are structured in a tabular format and are presented in a A3 format. All in all, six statements about the students´ abilities are listed, written in the first person singular ("I can…"). Each statement covers one subtopic of the chemical reaction. The subtopics "chemical reaction", "difference between chemical and physical reaction" and "chemical equation" are arranged together and the remaining subtopics "oxidation", "conservation of mass" and "chemical reaction with particles" are listed on the second self-evaluation-sheet. In order to identify what the students have learnt, the students asses themselves on a four-point Likert scale going from "I am very confident" to "I am not confident at all". After that they decide which proficiency levels they want to reach by using another four-point Likert Scale. Both, the assessment of their distinct achievement and the setting of a personal goal define the individual learning path, which the students pass independently. On the self-evaluation sheets the students find direct links to exercises in different levels of complexity and further informational texts. After completing an exercise, a feedback can be obtained by using sample solutions. Once students have finished a task, they document what material has been used.

Learning material

The learning material consists of informational texts, exercises and sample solutions which are used by the students during the self-regulated working phase. Between the two self-regulated working phases, the experimentation takes place. As additional guidance, the learners receive experimental instructions.

Informational texts

For each of the six subtopics, the learners are provided with informational texts on one A4 page, so that three explanations can be read during each of the self-regulated phases (90 minutes). Because of the fact that the lessons are an introduction to the topic of chemical reaction, it appeared reasonable to provide texts that summarized what previously was part of the presented short lectures. With regard to the UDL, especially the principles of the first guideline are implemented here since it focuses on the perceptual aspect of collecting

information. On this basis, important information is emphasized and information-supporting images are implemented. In addition to the visual perception, meaning independent reading, the following corresponding auditory variant of information is offered to the students: The learners have the option to read the informational texts by using a lecture-pen. The pen used in this intervention is the AnyBook reader from Franklin Discover. Informational texts were recorded by the researchers on the lecture pens prior to the lessons. Each class had five pens available during the intervention. For the preparation of the informational texts, the texts were laminated and customized according to the recordings on the memory sticks of the AnyBook reader. Short passages were chosen so that the students could read individual passages of the informational texts aloud. As the pen can recognise a specific code on stickers, these are put next to the written equivalent of the recorded auditory information linked to this code and to which the learners can listen to s with head phones. Of course, in addition to the laminated explanations, informational texts in the usual paper format are provided. Thus, each learner is able to decide for himself how he wants to access information.

Learning material

The six subtopics are represented by a three-stage differentiation. Thus, each subtopic includes three worksheets with different tasks. The cognitive demands on task management increase from simple to mediocre to challenging. Depending on the assessment and learning goal of the learners, the individual learning path is determined. For fast learners, there is an additional task at the end of a 90-minute lesson phase, which links content from three main areas. In total, nine worksheets of different difficulty level are made available to the students in each self-regulated learning phase, as well as a worksheet with linking tasks. In addition to the design aspects already described in relation to the informational texts, the differentiation into levels of complexity is another special factor of the UDL and is especially addressed within the third guideline regarding the promotion of persistent learning as this can be supported by different levels of challenge. It was particularly crucial in the chosen differentiation that there were three different worksheets with different types of tasks, each of which focus a common theme. A differentiation only in terms of the task seemed unsuitable for preventing the learners from working only on the quite simple exercise sheets.

Sample Solution

Especially when working independently, the feedback aspect of an activity should be given as much attention as possible which is why sample solutions are used to implement this element. Within the framework of the teaching unit, a sample solution is thus available to the learners for each worksheet which makes a total of nine sample solutions per 90-minute phase. Like all other developed materials, the sample solutions are based on the principles of the UDL which was especially taken into account in the design aspect. For example, the sample solutions also include pictures. In addition, the solutions are highlighted in different colours in order to make it easier for learners to see what the correct answer is. Furthermore, the use of sample solutions promotes self-regulated learning which is also part of the UDL. The sample solutions differed from the corresponding task sheets in their laminated form. The students are encouraged to use a red fibre pen when checking their results. In this way, we can later analyse later to what degree the students use the sample solution during the intervention.

Experimental Sheet

Altogether, five experiments were available for the pupils, which could be worked on independently by the learners with corresponding experimental instructions. This is because the experimental phase should also follow the principles of self-directed learning in order to satisfy the widest possible range of learners. At the beginning of the experimental phase, a short safety briefing covering the use of gas burners for example, was given. Since the pupils had little experience in experimenting, a special selection of experiments was required. In addition to the oral safety instruction, a poster in A0 format was also placed in the classroom, which presented all important safety-relevant aspects in text and pictorial form. As with the other materials of unity, the principles and guidelines of the UDL were also used in designing the experimental sheets to give as many learners as possible access to it. The presentation of the required materials as well as the execution steps were supported by photographs of the objects and actions. In addition, the students had the choice between recording their observation as a drawing or writing it on the experimental instructions. Common technical terms such as "execution" were supplemented by linguistically simplified descriptions such as "That's how you do it". As with the learning materials used in the self-regulated work phase, learners were able to control and correct their results with the help of sample solutions, using a red fibre pen again.

PARTICIPANTS

The participants in the main study were eighth-graders attending five secondary schools (*Gesamtschule*) in Germany (N = 224). Due to sickness related absences, the sample was reduced to n = 172 subjects (pre/post). Furthermore, data sets of 158 students could be used in the pre/follow-up data analysis.

MEASURING INSTRUMENTS

- Intellectual performance test: This instrument measures students' intellectual performance by doing one scale of the CFT 20 (Weiß, 1998) *before* the lessons.
- Self-concept scale: The second instrument assesses students' self-concept and is done *before* the lessons. It is adapted from DISK (Rost et al., 2007).
- Chemistry performance test: For this instrument, we developed a multiple-choice test consisting of 24 items with one correct answer out of five possible options. The test is done once *before* and twice *after* the lesson. The Cronbach alpha measure of internal consistency reliability for this test was .80.
- Feedback questionnaire: The fourth instrument was used *after* the lessons. It measures students' feelings towards the lessons. It contains 24 items. The five rating scale options range from *totally agree* to *totally disagree*. The Cronbach alpha measure of internal consistency reliability for this test was .89.
- Student skill assessment: The fifth instrument was used *before* the intervention. It measures students' skills using a rating by the teacher. It contains 16 items and a five-point Likert scale from *very good* to *not good at all*. The Cronbach alpha measure of internal consistency reliability for this sheet was .97.

RESULTS

We used data from the multiple-choice tests and the feedback questionnaire to find out whether there were differences regarding the learning progress among the groups.

Learning outcome

Due to the limited extend of this article, only those participants who have taken part in all three measurement periods of the study are taken into account below.

To determine possible differences, a residual analysis was done. Our results indicate significant learning outcomes in both groups from pre to post (Pre-Post: WithoutSEN $n = 87$, $p = <.001$, $\varphi = .84$; WithinSEN $n = 71$, $p = <.001$, $\varphi = .84$). The residual analysis shows no indication of a difference between the groups ($n = 158$, $p = .849$, $\varphi = .01$). Considering the long-term-effect, the learning outcomes also increased significantly from pre to follow-up (Pre-Follow-up: WithoutSEN: $n = 87$, $p = <.001$, $\varphi = .84$; WithinSEN: $n = 71$, $p = <.001$, $\varphi = .81$. A group comparison (pre-follow up) shows an almost significant difference in favour of the WithoutSEN Group ($n = 158$, $p = .053$, $\varphi = .15$). Since there has been no controlled intervention between the time of the post-measurement and the time of the follow-up, it is not possible to say which contents were dealt with after the intervention in the classroom.

Feedback

The students' feedback on the inclusive learning unit was positive ((Five-point Likert scale from 1 = *totally agree* to 5 = *totally disagree*) WithoutSEN $M = 2.25$; WithinSEN $M = 2.14$). There is no statistical difference between the groups ($n = 172$; $p = .253$; $\delta = 0.15$).

DISCUSSION AND CONCLUSION

The present study examines the question of whether the increase in learning in inclusive classes differs from that of non-inclusive classes. In a first step, a method containing both instructive and constructive elements was developed. The instructive part is represented by teacher presentations, while elements of the construction are covered by self-regulated learning. In order to ensure the best possible access to the content of the unit, the Universal Design for Learning was also implemented and especially taken into account when designing the learning materials.

The initial results show that there is no significant difference between the inclusive learning group (WithinSEN) and the non-inclusive learning group (WithoutSEN) groups in terms of both immediate and sustained knowledge growth. In addition, it can also be noted that both groups are equally positive about the teaching unit.

Since the intervention consists of three main elements, namely the instructive and constructive elements as well as the UDL, it cannot exactly be explained why the learners of both groups generate knowledge since the effect of the teacher's lecture or the self-evaluation sheet was not tested separately. This is due to the fact that there should be too much testing within the unit. Furthermore, it is also not clear in how far UDL lessons are more effective in comparison to conventional lessons. Overall, it can be assumed that the intervention has led to an increase in learning by combining the three central elements.

The study presented here was conducted under research conditions. It remains an open question to what extent the elements used can be transferred to teaching practices at schools. All in all, it must be taken into account that designing learning material based on the UDL is time-consuming. On the other hand, however, UDL lessons carry extra value for the students. Lastly, having appropriate materials for inclusive teaching can contribute to reduce the overall burden on teachers in schools.

ACKNOWLEDGEMENT

Finally, I would like to thank my entire working group and all students and teachers, who participated in the project, for their cooperation.

REFERENCES

Center of Applied Special Technology (CAST) (2011). Universal Design for Learning Guidelines 2.0. Wakefield, MA: Author

Convention on the Rights of Persons with Disabilities (CRPD) (2006). www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html (02.01.2018)

Engelmann, S. (1980). The instructional design library. New Jersey: Educational Technology Publications Englewood Cliffs.

Grell, J. (1999). Direktes Unterrichten: Ein umstrittenes Unterrichtsmodell. In J. Wiechmann (Ed.), Zwölf Unterrichtsmethoden. Vielfalt für die Praxis (pp. 35–49). Weinheim, Basel: Beltz.

Gruehn, S. (2000). Unterricht und schulisches Lernen: Schüler als Quelle der Unterrichtsbeschreibung. Münster: Waxmann.

Hasselhorn, M., & Gold, A. (2009). Pädagogische Psychologie: erfolgreiches Lernen und Lehren. Stuttgart: Kohlhammer.

Klahr, D., Nigam, M. (2004): The Equivalence of Learning Paths in Early Science Instruction: Effects of Direct Instruction and Discovery Learning. *Psychological Sience*, *15*(10), 661-667.

Kullmann, H., Lütje-Klose, B., & Textor, A. (2014). Eine Allgemeine Didaktik für inklusive Lerngruppen – fünf Leitprinzipien als Grundlage eines Bielefelder Ansatzes der inklusiven Didaktik. In: B. Amrhein & M. Dziak-Mahler (Eds.):*Fachdidaktik inklusiv*. (p. 96-97). Münster: Waxmann.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (2009). Erstes Gesetz zur Umsetzung der UN-Behindertenrechtskonvention in den Schulen (9. Schulrechtsänderungsgesetz). www.schulministerium.nrw.de/docs/Schulsystem/Inklusion/Gesetzentwurf.pdf (02.01.2018).

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33-40.

Quittenbaum, N. (2016). Training für direkte Instruktion: Die Entwicklung und Erprobung eines Kommunikationstrainings für den Unterricht mit direkter Instruktion. Bad Heilbrunn: Julius Klinkhardt.

Rost, D. H., Sparfeldt, J. R., & Schilling, S. R. (2007). DISK-Gitter mit SKSLF-8. Differentielles Schulisches Selbstkonzept-Gitter mit Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten. Göttingen: Hogrefe.

UNESCO. (1994). The Salamanca statement and framework for action on special needs education. www.right-to-education.org/resource/salamanca-statement-and-framework-action-special-needs-education (02.01.2018)

Weiß, R. H. (1998). Grundintelligenztest Skala 1 (CFT 20). Göttingen: Hogrefe.

Tuovinen, J. E., & Sweller, J. (1999): A Comparison of Cognitive Load associated with discovery learning and worked examples. *Journal of Educational Psychology*, *91*(2), 334-341.

Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology*, *80*(3), 284-290.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

REPRESENTATIONS OF THE PCK BEFORE AND AFTER THE SUMMIT

Brunno Carvalho Gastaldo[1]; Pablo Micael Castro[2]; Paula Homem-de-Mello[1] and Sérgio Henrique Leal
[1]Federal University of ABC, Brazil
[2]University of São Paulo, Brazil

Pedagogical Content Knowledge (PCK), proposed by Shulman, has occupied the centre of effort to capture teachers' expertise. Since Shulman´s work, several models have been developed, in order to better define the PCK. However, some of these derivations have diminished the strengths of the construct, because they can disagree in important aspects. To resolve this, a congress called PCK Summit was held, wherein the PCK Consensus Model was developed. To analyse if PCK Summit influences PCK research, this paper focuses on understating PCK Summit's effect on PCK representations using a lexicometric exploratory, descriptive, and comparative analysis. We employed Descending Hierarchical Classification (DHC) and Factorial Analysis of Correspondence (FAC) of text segments, extracted from the corpus composed of papers containing the acronym "PCK", published before (α), during (ε) and after (β) the PCK Summit. Our results show that the papers published during and after PCK Summit have a more mature view of the PCK than the α papers, and are about quantitative methods and curricula modifications to PCK development. Moreover, the representations found in β papers have an intense relation with the PCK Summit's work groups, and that the topic-specificity of the PCK, has gained more attention in β works, appearing in two discourses. Therefore, it is possible to conclude that the PCK Summit has influenced PCK representation.

Keywords: descending hierarchical classification (DHC); factorial analysis of correspondence (FAC); PCK summit

INTRODUCTION

The desire to capture teacher's expertise is ancient, and there is no final agreement on the qualities of a good teacher (Barnhart & van Es, 2015). However, since Shulman's work (Shulman, 1986), the Pedagogical Content Knowledge (PCK) has occupied the centre of such effort, as it is said to comprehend the body of knowledge needed for teaching. From then on, several models were developed trying to better define the PCK. This rich effusion of propositions allowed a vast array of research covering different aspects and contexts. On the other hand, it diminished the strengths of the construct, once researchers diverged on what is exactly the PCK and its components, leading to the participants noticing an increased difficulty in publishing PCK articles (Borowski et al., 2011).

To mitigate the existing disagreement on the used vocabulary and the nature of the PCK (personal or canonical), ways to assess / measure it and its topic or domain specificity, a "congress like event" (Helms & Stokes, 2013) was held in Colorado Springs in 2012, gathering researchers from 13 research groups (Table 1) with the objective of exploring "the potential of a consensus model of PCK to guide science education research [and the] identification of

specific next steps [to] move the field forward" (Carlson, Stokes, Helms, Gess-Newsome, & Gardner, 2015, p. 16).

Table 1. Researchers attending the Summit and their research group.

G1	G2	G3	G4	G5	G6	G7
K. R. Daehler, J. I. Heller, J. W. Little, K. Sheingold, M. Shinohara, N. Wong	J. Gess - Newsome, J. Carlson, A. Gardner	R. Schneider	A. Berry, R. Cooper, J. Loughran	M. Rollnick, E. Mavhunga	E. Banilower, S. Smith	J. van Driel, I. Henze
G8	G9	G10	G11	G12	G13	
P. Friedrichsen, J. Lannin, A. Sickel	V. Kind	K. Padilla, A. Garritz	H. Hill, D. L. Ball, H. Bass, M. Blunk, M. Thames, J. Lewis, G. Phelps, L. Sleep	S. Kirschner, A. Borowski, H. Fischer	S. Park, J. K. Suh	

In preparation to the event, the organizers took some precautions in order to enrichen the debate. The most important, according to the participants, was writing a conference paper detailing their PCK research program (e.g. their definition of PCK, model used, assessment tools, etc.) and also reading thoroughly theirs peer papers (Helms & Stokes, 2013).

Through the days of the event, forums (Table 2) were held allowing them to share their different views in small groups to solve discrepancies, and then in large ones to share the conclusions. At the final days participants were encouraged to form Work Groups (WG) according to the emerging interests (BSCS, 2012) Table 3).

Table 2. Forums held in the first days the Summit (BSCS, 2012).

	Forums	Groups
1	Content Knowledge and PCK	G2, G11, G12
2	Beliefs, Teaching Orientation, and PCK	G8, G9, G10
3	Nature of PCK	G4, G6, G7
4	PCK Models and Assessment Implications	G5, G8 G12,
5	Assessment of PCK	G4 G6, G13,
6	Research Findings on PCK	G1, G2 G3,

Table 3. Work Groups (WG) held at the final days of the Summit (BSCS, 2012).

	Work Group
WG1	Refining the PCK model
WG2	Developing PCK in teachers (over the trajectory from pre-service to experts)
WG3	The research map for PCK
WG4	Connecting PCK to policy

In the last day of the event, a model was developed and named Consensus Model of PCK, and both Canonical and Personal PCK were defined. The former is the one that can be shared and is substantiated by systematic research (Rollnick & Mavhunga, 2015), whereas the latter is "the knowledge of, reasoning behind, and enactment of the teaching of particular topics in a particular way with particular students for particular reasons for enhanced student outcomes" (Garritz, 2015; Helms & Stokes, 2013).

Conversely, five years have passed, the participants seem to keep investigating in their specific fields of interest, and above all, not using the Consensus Model of PCK.

In this work, we aim to evaluate whether the Summit has affected or not the participants research, and if so, how those changes appear in their representation of the PCK in their latter papers. To do so, an optimal way is performing a lexicometric analysis as it "enables extracting the pattern of social representations of an object from *corpora* in natural language" (Lahlou, 1996, p. 279), and especially using Computer Assisted Qualitative Data Analysis (CAQDA) as it enables mining information from large *corpora* (Costa, Reis, Sousa, Moreira, & Lamas, 2017). It is also vastly used in the educational research field, above all to understand the interactions between students and teacher (e.g. Lewins & Silver, 2007; Mortimer & Scott, 2002; Sickel, Witzig, Vanmali, & Abell, 2013), nonetheless, is has a scarce usage in scientific texts (Atanassova, Marc, & Mayr, 2015; Bertin & Atanassova, 2015).

The lexicometric analysis, first proposed by Lebart & Salem(1988), was formalized in a software (Alceste®) by Reinert (Reinert, 1990), allowing an increase in the *corpus* size. In this paper, an open code version of the software developed by Ratinauld was used (Lowen, Peres, Crozeta, Bernardino, & Beck, 2015; Ratinaud & Marchand, 2012). The software divides the corpus in text segments and compares the frequency of their words in each segment, and then classifies the text segments with similar words together using a chi-squared (χ^2) test (Camargo, 2005). Those classes show the different types of discourse present in the text, as "meaning may be studied through the way people use words in combination with other words" (Chartier & Meuneier, 2011, p. 8; Garnier & Guérin-Pace, 2010; Lahlou, 1996; Sommer Harrits, 2011).

Therefore, this paper focused on understanding if, and how, the representations of PCK changed after the Summit, and also to establish if the Summit can be inferred an INUS condition (insufficient but non-redundant parts of a condition which is itself unnecessary but sufficient for the occurrence of the effect)(Mackie, 1965).

METHODS

In order to understand the changes in representations of PCK, if any, in the production of the participants before and after the PCK Summit, a lexicometric exploratory, descriptive, and comparative analysis was performed.

First, the papers from 5 years before (α), the conference papers (ε), and 5 years after (β) the Summit were collected from the data bases: Google Scholar, Research Gate, ERIC and Directory of Open Access Journals (Harzing & van der Wal, 2008; Meho & Yang, 2007) and used in full to form a *corpus*. Other restrictions were: being written in English, being peer-reviewed, and having at least one author attending the Summit. They were normalized from

idiom variances and terminology used, and then, for this analysis, a *sub-corpus* was created with the text segments containing the acronym "PCK" originating the text segments *πα*, *πε* and *πβ*, respectively.

The analysis was performed in the software IRAMUTEQ®, and the text segments contained 40 words and 12 tokens text segments vs. 14 tokens, with a maximum of 10 classes (standard parameter) (Gobbo & Same, 2016) and with lemmatization (Sarrica, Mingo, Mazzara, & Leone, 2016). Utilizing that *sub-corpus*, a Descending Hierarchical Classification (DHC) was developed and Factorial Analysis of Correspondence (FAC) was performed (Chartier & Meuneier, 2011; Costa et al., 2017; Lahlou, 1996). The DHC analysis was made regarding the 10 words with higher chi^2 (χ^2) in the class, which enables the recognition of the typical features tagging it by its synthase semantic content, in an hermeneutical analysis (Chartier & Meuneier, 2011; Lahlou, 1996, 2012); and also looking at the most significant text segments (containing a higher sum (Σ) of χ^2 from the its' words).

To increase trustworthiness, all data was analysed by two independent researchers and the methods and data were deposited in the Center for Open Science's Open Science Framework to assure transparency (Gastaldo & Castro, 2017).

RESULTS AND DISCUSSION

The analysis has shown that, the papers published before the Summit (α) showed two different representations about the PCK, the first (α - 1/2) was named *PCK Model* and focused on understanding how the PCK is constructed. Almost all groups who were related to any specific discourse fit into this class: they were groups 5, 8, 10, 13 (Figure 1). The words with a higher χ^2 were: *component, model, knowledge, Magnusson, SMK, orientation, Shulman, belief, include* and *category*. As for the second discourse (α - 2/2), which was named *Development of PCK and CoRe*, there is a marked presence of the CoRe instrument, both as a mean to assess the PCK and develop it. Although it has more than half of the text segments from the papers before the Summit, only one group contributed to it, group 4. The words with a higher χ^2 were: *CoRe, student, development, participants, PaP-eRs, learn, preservice, practice, construct* and *educator*.

It is important to acknowledge that this was the group that developed this instrument, but, despite it being vastly used throughout literature, its presence is such that it establishes a distinct discourse.

The almost homogeneity of the discourses amongst those researchers can be justified by their interest in understanding the nature of the PCK and its' origins in the teacher formation. Many papers discuss how the PCK is originated and what are its components (e.g. Berry, Loughran, & van Driel, 2008; Garritz, 2010; Henze, van Driel, & Verloop, 2008; Nilsson & van Driel, 2011).

This homogeneity is broken in *πε*, where it is possible to see five different discourses. They show that researchers wandered in many directions trying to characterize the PCK, and it reflected in the way that PCK is represented.

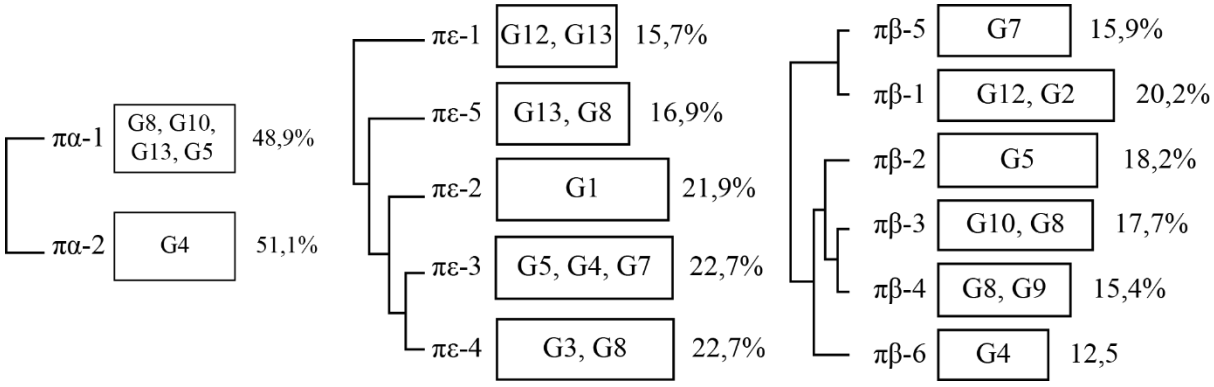


Figure 1. PCK representations found in the three moments. The groups in each discourse are presented in order by their contribution. Those within a double circle contributed tree times more than the following and the ones with a single circle twice as much as the following group.

The first distinct discourse is *πε – 1* which was named *Quantitative analysis of PCK*. It inaugurates not the usage of quantitative methodologies to investigate the PCK, but a quantitative discourse to represent the PCK, having as responsible for it groups 12 & 13. The words with a higher χ^2 were: *CK, PK, dimension, physic, validity, professional knowledge, distinct, test, correlation,* and *task*

The second discourse found (*πε – 2*) was related to the *Relation between PCK and students*, being interested in teachers' knowledge of students understanding of Subject Matter and their way of thinking. In this discourse class, text segments from a wide number of groups can be found, being G1 the only one who was statistically related to it. The words with a higher χ^2 were: *student, learn, specific, SMK, concept, notion, relate, understand, content,* and *lead*.

Discourse number 3 (*πε-3*) was named *Teacher profession development research context* as it has a marked desire to understand the researches related context in which the teachers develop their PCK. The groups and words more strongly associated to this discourse are: G4, 5 & 7 and *preservice, investigate, in service, context, science, group, study, instance, validation,* and *educational*.

As for the fourth discourse found (*πε – 4*), the main idea behind its text segments was the *Teacher profession development programs.* As it can be seen on Figure 1, *πε – 4* and *πε – 3* are closely related, and both are dedicated to the representation of the environment of PCK development. However, in this case, the representation is not focused on the researches but on the development of programs itself. The main groups that produced this discourse were G3 & G8 whose main words were *program, design, support, professional development, research, education preparation, year, educative,* and *course*.

The last discourse (*πε – 5*) opposes the 3 predecessors (*πε – 2*, *πε – 3* & *πε – 4*) as, alongside with *πε-1*, it does not represent the PCK development, but the *PCK Models & components*. Nonetheless, it shares similarities with the same 3 as they are all related to theoretical aspects of the construct, in contraposition to *πε – 1* which representation, as already discussed, is dedicated to methodological aspects. The groups and words with higher χ^2 were: G8 & G13 and *component, Borko, description, distinct, Mulhall, Berry, KISR (Knowledge of Instruction Strategies and Representations), Krajcik, pentagon,* and *Grossman.*

Regarding the papers published after the Summit (*πβ*), a wider matrix can be observed as six different discourses were found (Figure 1).

The first discourse (*πβ* – 1) follows the tendency of *πε* – 1 and is concerned with the *Quantitative theory of PCK*. That being said, in this class, PCK representations are particularly connected with specificities of the quantitative analysis and methodology, which can be seen in the words with higher χ^2: *Content Knowledge, Pedagogical Knowledge, dimension, physic, validity, professional knowledge, distinct, test, correlation,* and *task*. As for the most important groups of this class, G12 and G2 can be pointed out.

The second discourse (*πβ* – 2) was named *Topic specific level of PCK & SMK transformation*, and a has a direct relationship with the Summit. It can be said that the topic specificity of PCK is not a new idea. However, the Summit grants a validation that makes it possible for G5 to create a body of text segments explicitly related to it, and by that means, allowing it to be identified as a distinct discourse. Such analysis is strengthened by the excerpt

> Like our models the version of the model emerging from the summit separates teacher knowledge domains from a construct referred to as topic specific professional knowledge which aligns to our TSPCK (Rollnick & Mavhunga, 2015)

Amongst others which even evoke Shulman as an *argumentum ad antiquitatem* (e.g. Mavhunga, Ibrahim, Qhobela, & Rollnick, 2016; Mavhunga & Rollnick, 2013). The words with high χ^2 that were used in this analysis alongside with the text segments (not shown) were *Makinster, level, Veal, Shulman, transformation, equilibrium, specific, chemical, programe,* and *context*.

The third discourse (*πβ* – 3) focuses on the *Prospective curricular changes to increase the PCK* and corresponds to the still existing gaps within the PCK field and points to ways to amend them. The groups that produced the text segments creating this PCK representation were G8 & G10, and they used as main words: *evolution, curriculum, rich, kind, program, science, SMK, teacher, education,* and *understanding*.

The *topic specific level of PCK & PCK components* were the subjects addressed by the forth discourse (*πβ* – 4). The PCK components theme returns in this discourse yet as a consolidated feature of PCK, not in the exploratory version as before. Groups G8 & G9 are the ones with significant relations to this discourse, and the main words are: *component, topic, purpose, orientation, Magnusson, specific, Friedrischen, Science Teaching Orientations,* and *compare*.

In the same cluster as *πβ* – 1, *πβ* – 5 represents PCK by means of *quantitative measurements*, and as the former addresses epistemological aspects, the latter deals with more practical characteristics. With group 7 as the characteristic one, the most relevant words were: *item, test, sample, score, biology, scale, objective, evaluation, open,* and *main*.

The last discourse from *πβ* (*πβ* – 6 – *CoRe use for portraying PCK*) is the most diverse, apart from the quantitative super-class (classes 1 & 5). As before with *πα* – 2 & *πε* – 4, this class is highly associated with Loughran's research, that continued producing a PCK representation that relates to the CoRe instrument to assess and develop the PCK, at a point which this last type of discourse is exclusively produced by his group 4 (as is in its' origins in *πα* – 2). The

main words were: *CoRe, associate, student, phase, practicum, process, interview, Hume, source,* and *prompt*.

Reckoning this data, is possible to affirm that the PCK representation matrix has a crescent complexity and that, after the Summit, there are more different and vast topics related to PCK representation, as Friedrichsen affirms: "questions have increased in number and […] in refinement" (2015, p. 159).

One of the changes observed was the rise of the quantitative representation after its first appearance during the Summit, and even the groups that do not present a quantitative representation use quantitative methodology.

On the other side of the scale, a decrease of papers regarding the PCK models is evident. As the forums' themes had the goal of solving issues intriguing the researchers until that time, one could predict that those themes would disappear from the PCK representation, reaching a more mature version. This predicable phenomenon, in truth, happened in *πβ* particularly with the themes motivating the forums 2 – *Beliefs, Teaching Orientation, and PCK* & 3 – *Nature of PCK*, which do not relate to any *πβ* representation.

Finally, the representations of PCK gain a new feature in πβ. Those representations have a strong relation with the work groups (WG) held in the Summit. There is a clear semantical relation between WG1 – Refining the PCK model and representations *πβ* – 1 – *Quantitative theory of PCK* & *πβ* – 5 – *Quantitative measurement of PCK*, which expand the way to represent the PCK particularly as they present themselves as new representations and thus in more need to be expanded.

Close relations are also found between WG2 and *πβ* – 2 – *Topic specific level of PCK & SMK transformation*, πβ – 3 – *Prospective curricular changes to increase the PCK*, and *πβ* – 4 – *Topic specific level of PCK & PCK components*. Even more direct is the relation of the third WG – *Connecting PCK to policy* and *πβ* – 3- *Prospective curricular changes to increase the PCK*.

CONCLUSION

Our results show that the PCK Summit had a noticeable impact in PCK representations on the attending authors, even though they still do not use the Consensus Model, nor do they officially employ the new PCK definition. The papers presented at such encounter have a more mature view of PCK, and, after it, they showed the discussions made in it. The quantitative discourse appears on *πε* and pervades *πβ*, refining PCK representations and establishing a temporal precedence.

The Work Groups held at the end of the Summit have a close semantic relation with the PCK representations of the papers published after the Summit, indicating that, although the Consensus Model is not being adopted as a heuristic tool, the cognitive work developed at such event influenced the way the PCK is addressed.

By this effect, it is possible to affirm that the Summit constitutes an INUS condition, as it contributed non redundantly to what those researchers produced after it.

REFERENCES

Atanassova, I., Marc, B., & Mayr, P. (2015). Mining Scientific Papers for Bibliometrics: a (very) Brief Survey of Methods and Tools. *15th International Society of Scientometrics and Infometrics Conference*.

Barnhart, T., & van Es, E. (2015). Studying teacher noticing: Examining the relationship among pre-service science teachers' ability to attend, analyze and respond to student thinking. *Teaching and Teacher Education*, *45*, 83–93. https://doi.org/10.1016/j.tate.2014.09.005

Berry, A., Loughran, J., & van Driel, J. H. (2008). Revisiting the Roots of Pedagogical Content Knowledge. *International Journal of Science Education*, *30*(10), 1271–1279. https://doi.org/10.1080/09500690801998885

Bertin, M., & Atanassova, I. (2015). Factorial correspondence analysis applied to citation contexts. *CEUR Workshop Proceedings*, *1344*, 22–29.

Borowski, A., Carlson, J., Fischer, H. E., Henze, I., Gess-Newsome, J., Kirschner, S., & van Driel, J. (2011). Different Models and Methods to Measure Teachers' Pedagogical Content Knowledge. In *ESERA 2011 Conference*.

BSCS. (2012). PCK Summit Dissemination Site. Retrieved from http://pcksummit.bscs.org/node/81

Camargo, B. V. (2005). ALCESTE: Um programa informático de analise quantitativa de dadostextuais [ALCESTE: A computer program for quantitative analysis of textual data]. In A. S. P. Moreira, B. Camargo, J. C. Jesuino, & S. Nóbraga M (Eds.), *Perspectivas teórico-metodológicas em representações sociais [Theoretical-methodological perspectives in social representations]* (pp. 511–539). João Pessoa, Paraíba, Brazil: Editora Universitária UFPB.

Carlson, J., Stokes, L., Helms, J., Gess-Newsome, J., & Gardner, A. L. (2015). The PCK Summit: A process and structure for challenging current ideas, provoking future work, and considering new directions. In *Re-examining Pedagogical Content Knowledge in Science Education* (pp. 14–27). New York: Routledge.

Chartier, J. F., & Meuneier, J. G. (2011). Text Mining Methods for Social Representation Analysis in Large Corpora. *Papers on Social Representation*, *20*, 37.1-37.47.

Costa, A. P., Reis, L. P., Sousa, F. N. de, Moreira, A., & Lamas, D. (2017). *Computer Supported Qualitative Research*. (J. Kacprzyk, Ed.), *Studies in Systems, Decision and Control*. Switzerland: Springer.

Friedrichsen, P. M. (2015). My PCK research trajectory: A purple book prompts new questions. *Research in Science Education*.

Garnier, B., & Guérin-Pace, F. (2010). *Appliquer les méthodes de la statistique textuelle [Apply the methods of textual statistics]*. Retrieved from http://www.ceped.org/?Appliquer-les-methodes-de-la

Garritz, A. (2010). Pedagogical Content Knowledge and the Affective domain of Scholarship of Teaching and Learning. *International Journal for the Scholarship of Teaching and Learnind*, *4*(2).

Garritz, A. (2015). PCK for dummies. Part 2: Personal vs Canonical PCK. *Educación Química*, *26*(2), 77–80. https://doi.org/10.1016/j.eq.2015.04.001

Gastaldo, B. C., & Castro, P. M. A. (2017). The Summit effect on PCK representation. Retrieved from http://osf.io/3dmzr

Gobbo, A., & Same, F. (2016). 20 years of OK budget speeches: correspondence analysis vc. networks of n-grams. In *13 Journées internationales d'analyses stetistique des donnés textualles*.

Harzing, A. W. K., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*, 61–73. https://doi.org/10.3354/esep00076

Helms, J., & Stokes, L. (2013). *A Meeting of Minds around Pedagogical Content Knowledge: Designing an International PCK Summit for Professional, Community, and Field Development*. *PCK Summit*. Inverness Research.

Henze, I., van Driel, J. H., & Verloop, N. (2008). Development of experienced science teachers' pedagogical content knowledge of models of the solar system and the universe. *International Journal of Science Education*, *30*(10), 1321–1342. https://doi.org/10.1080/09500690802187017

Lahlou, S. (1996). A method to extract social representations from linguistic corpora. *Japanese Journal of Experimental Social Psychology*, *35*(3), 278–291. https://doi.org/10.2130/jjesp.35.278

Lebart, L., Salem, A., & Dunod. (1988). *Analyse statistique des données textuelles. Questions ouvertes et lexicométrie [Statistical analysis of textual data. Open questions and lexicometry]*. (Dunod, Ed.). Retrieved from https://issuu.com/sfleury/docs/st-1994-lebart_salem/359

Lewins, A., & Silver, C. (2007). *Using software in qualitative research: a step-by-step guide*. Great Britain: SAGE Publications.

Lowen, I. M. V., Peres, A. M., Crozeta, K., Bernardino, E., & Beck, C. L. C. (2015). Managerial nursing competencies in the expansion of the family health strategy. *Revista Da Escola de Enfermagem*, *49*(6), 964–970. https://doi.org/10.1590/S0080-623420150000600013

Mackie, J. L. (1965). Causes and Conditions. *American Philosophical Quarterly*, *2*(4), 245–264. Retrieved from http://www.jstor.org/stable/20009173 .

Mavhunga, E., Ibrahim, B., Qhobela, M. M., & Rollnick, M. (2016). Student teachers' competence to transfer strategies for developing PCK for electric circuits to another physical sciences topic. *African Journal of Research in Mathematics, Science and Technology Education*, *20*(3), 299–313. https://doi.org/10.1080/18117295.2016.1237000

Mavhunga, E., & Rollnick, M. (2013). Improving PCK of Chemical Equilibrium in Pre-service Teachers. *African Journal of Research in Mathematics, Science and Technology Education*, *17*(March 2015), 113–125. https://doi.org/10.1080/10288457.2013.828406

Meho, L., & Yang, K. (2007). Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science Versus Scopus and Google Scholar. *Jounal of the American Society for Infermation Science and Technology*, *58*(13), 2105–2125. https://doi.org/10.1002/asi.2067710.1002/asi

Mortimer, E. F., & Scott, P. (2002). Atividade discursiva nas salas de aula: uma ferramenta sociocultural para analisar e planejar o ensino [Discursive activity in classrooms: a sociocultural tool to analyze and plan teaching]. *Investigações Em Ensino de Ciências*, *7*(3), 282–306. Retrieved from http://www.diaadiaeducacao.pr.gov.br/diaadia/diadia/arquivos/File/conteudo/artigos_teses/Ciencias/Artigos/mortimer_scott.pdf

Nilsson, P., & van Driel, J. (2011). How Will We Understand What We Teach? - Primary Student Teachers' Perceptions of their Development of Knowledge and Attitudes Towards Physics. *Research in Science Education*, *41*(4), 541–560. https://doi.org/10.1007/s11165-010-9179-0

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE aux "gros" corpus et stabilité des "mondes lexicaux": analyse du "CableGate" avec IRAMUTEQ [Application of the ALCESTE method to "big" corpora and stability of "lexical worlds": analysis of "CableGate" with IRAMUTEQ]. *Actes Des 11èmes Journées Internationales d'Analyse Des Données Textuelles (JADT)*, 835–844.

Reinert, M. (1990). Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval [Alceste a methodology for analyzing textual data and an application: Aurelia De Gerard De Nerval]. *Bulletin de Méthodologie Sociologique*, *26*(1), 24–54. https://doi.org/10.1177/075910639002600103

Rollnick, M., & Mavhunga, E. (2015). The PCK summit and its effect on work in South Africa. In *Re-examining Pedagogical Content Knowledge in Science Education* (pp. 135–146).

Sarrica, M., Mingo, I., Mazzara, B., & Leone, Gi. (2016). The effects of lemmatization on textual analysis conducted with IRaMuTeQ: results in comparison. In *13ème Journées internationales d'Analyse statistique des Données Textuelles*. Nice: Université de Nice Sophia Antipolis.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Sickel, A. J., Witzig, S. B., Vanmali, B. H., & Abell, S. K. (2013). The Nature of Discourse throughout 5E Lessons in a Large Enrolment College Biology Course. *Research in Science Education*, *43*(2), 637–665. https://doi.org/10.1007/s11165-012-9281-6

Sommer Harrits, G. (2011). More Than Method?: A Discussion of Paradigm Differences Within Mixed Methods Research. *Journal of Mixed Methods Research*, *5*(2), 150–166. https://doi.org/10.1177/1558689811402506

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

A COMPARISON OF STUDENT RESPONSES TO PICTORIAL AND VERBAL ITEMS FOCUSING ON CONCEPTUAL UNDERSTANDING OF THE PARTICLE MODEL OF MATTER

Elon Langbeheim[1], Emine Adadan [2], Sevil Akaygun[2], Manzini Hlatswayo[3] and Umesh Ramnarain[3]
[1]Department of Science Teaching, Weizmann Institute of Science, Rehovot, Israel
[2] Bogazici University, Istanbul, Turkey
[3]University of Johannesburg, Johannesburg, South Africa

We show that student reasoning about the particle model of matter is sensitive to pictorial and verbal formats of conceptual questions. This phenomenon is consistent across ages and curricula although the magnitude of the differences varies. We used a randomized trial in which a pictorial and verbal format of the same questions were assigned to students in the same classrooms. We administered the same questionnaire to three groups of secondary school students in three countries and found a significant difference in student response patterns between questions formats across all three groups. We suggest a more nuanced approach to the analysis of student ideas about matter that combines verbal and pictorial cues. Such an approach might have important implications on the design of curricula and learning progressions concerning the particle model of matter.

Keywords: mental models, visualizations, misconceptions

INTRODUCTION

The literature concerning student ideas about matter is dominated by Piagetian approaches that assign students to "stages" based on their responses to survey or interview questions (e.g. Johnson, 1998; Merrit & Krajcik, 2013; Hadenfeldt et al., 2016). For example, Johnson (1998) suggested that students' ideas about matter may be characterized along four general mental models and that student understanding can be described as progressing from one model to the other. According to Johnnson (1998), students holding naïve models of matter do not think of matter in terms of particles at all, students with a slightly more developed mental model, know that matter contains particles, but view particles as additional to the substance that comprises matter. Students at yet a higher level, think of matter as made of particles, but imagine these particles as having the same appearance and characteristics as the observable, macroscopic piece of matter. Students that acquire the "scientifically accurate" model, think of matter as comprised of particles, and understand that the properties of matter are collective, and that often, macroscopic observations of matter do not resemble the particle level behavior and appearance.

Developing such taxonomies of student ideas is important for science education, because it gives educators tools to identify the thinking of their students and to address it. However, such simplified categorizations of "flawed" understanding, may also overlook the important intuitive ideas that students use when learning these topics. Studies that challenge the mental models approach have shown that students' misconceived reasoning encompasses the

activation of many fruitful knowledge pieces (e.g. Smith, et al., 1993). The activation of certain ideas depends on how students frame the problem (Hammer, Elby, Scherr & Redish, 2005) and how they interpret the information embedded in it (Langbeheim, 2015). For example, a problem entailing a familiar context such as a person driving in a car, may be framed by the student as one that calls for "everyday" reasoning, and not the Newtonian principles of force diagrams and equations that were discussed in the classroom context. In the case of the particle model of matter, problems that contain illustrations may be elicit different sets of ideas than verbal information and direct students towards different conceptualizations.

The current study re-examines the use of "reasoning levels" or mental model levels to characterize students' conceptions of matter (Merrit & Krajcik, 2013; Hadenfeldt, et al., 2016). We explore whether information presented in a picture rather than in a written text, primes different patterns of reasoning about the structure of matter and physical processes in matter. Although researchers have recommended using both drawings and written text for eliciting student ideas about matter (Nussbaum & Novick, 1982), studies of student-made drawings of models (Nyachwaya et al., 2011) vis-à-vis student-made written descriptions of the same process are scarce. One such pioneering study (Akaygun & Jones, 2014) compared the prominence of structural and dynamic features in student self-generated particle models in drawings and in written explanations. In order to examine how does the presentation of information affect students' reasoning patterns, we compared the reactions to pictorial and verbal survey items in three groups of learners of different age levels, who are exposed to different curricula.

METHOD

Two forms of a questionnaire containing eleven equivalent pictorial and verbal questions adapted from a prior study (Hadenfeldt et al., 2016) were randomly administered to secondary school students in South Africa (10th grade, N=126), Israel (7th grade, N=90), and Turkey (7th grade and 10th grade, N=90). The schools in Turkey and Israel were semi-private schools with enrollment of students above the national average, whereas the South-African school was a public school with student level reflecting the national average. Three sample items from these questionnaires are shown in Figures 1-3. The item in Figure 1 elicits student ideas about the structure of water, and the item in Figure 2 elicits ideas about the configuration of gas particles and the item in Figure 3 elicits ideas about the process of dissolving sugar.

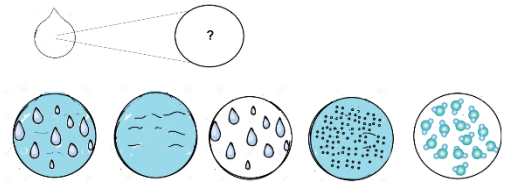
In all three items, some of the choices where designed to represent the less "sophisticated" reasoning levels or mental models. For example, in item 1, options A and C, represent water particles as resembling the shape of the macroscopic water droplets, and thus a flawed or incomplete model in which the molecular-level entities maintain the form of the macroscopic liquid. Option A and D illustrate a "hybrid" conceptualization of matter, in which particles are perceived as embedded in the macroscopic liquid, but not as the building blocks of liquid itself.

Similarly, in item 5 shown in Figure 2, air is released from a balloon. In this case students might think that the lower part of the balloon would be empty and choose "The remaining air particles stay at the top of the balloon" (option C), whereas the appropriate response would be " The remaining air particles scatter evenly throughout the balloon" (option D).

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

Item 1. Julie wants to sleep but the dripping faucet in the bathroom in the room next door keeps her up. While she lies in her bed, she imagines how water is composed. How do you think the particles of which water is composed look like?

Pictorial Format:



A B C D E

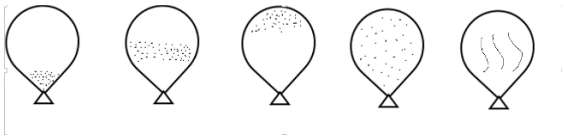
Verbal Format:

A. The water contains many particles that look like water drops.
B. There are no particles in the water drops
C. The water particles look like water drops and are surrounded by air.
D. The water particles look like small balls that swim in water.
E. There are many tiny water particles in a drop of water that do not look like droplets.

Figure 1. Water droplet item: eliciting student ideas about the internal structure of matter

Item 5. Some air is released from a balloon. The balloon is closed afterwards.
How do the remaining air particles arrange in the balloon?

Pictorial Format:



A B C D E

Verbal Format:

A. The remaining air particles bunch-up near the balloon's knot.
B. The remaining air particles bunch-up in the middle of the balloon.
C. The remaining air particles stay at the top of the balloon.
D. The remaining air particles scatter evenly throughout the balloon.
E. The remaining air fills the entire balloon

Figure 2. Spreading of air in a balloon eliciting ideas about the spatial configuration of gas particles

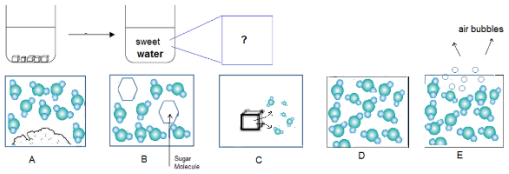
A final example is shown in Figure 3. The macroscopic disappearance of the sugar in the water, corresponds to answers such as "the sugar turns into water" (option C), or "the sugar particles disappear but leave a sweet taste behind" (option D).

The questionnaire was designed as a two-tier questionnaire, in which students first answered the multiple-choice item, and were then asked to explain their choices. The required explanation in the second tier was verbal if the item choices were pictorial, or pictorial if the item choices were verbal. Students' drawings and written explanations were analyzed in order to compare their self-produced representations and the equivalent verbal/pictorial representation that was used in the multiple-choice options.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

Item 8: When we add a sugar cube to hot water and stir, the sugar cube is no longer seen. Which of the following explains what happens to the sugar when it is added to the hot water?

Pictorial format:



Verbal Format:

A. The sugar scatters to the bottom of the cup
B. The sugar dissolves, and the sugar particles mix with the water particles
C. The sugar particles became water particles
D. The sugar disappears, and only the sweet taste is transferred to the water
E. The sugar particles become air particles that form bubbles at the surface of the water and then escape from the water

Figure 3. Sugar dissolved in hot water - eliciting ideas about the process of mixing

FINDINGS

An item-by-item comparison of student responses revealed differences in the response patterns, which were consistent across all three groups. In six out of the eleven items, the proportions of responses to the pictorial and verbal items were similar, whereas in five of the eleven items we found significant differences in the response patterns to the verbal and pictorial formats. Such significant differences are illustrated in the response patterns illustrated in Figures 4 & 6, similar response rates are illustrated in Figure 5.

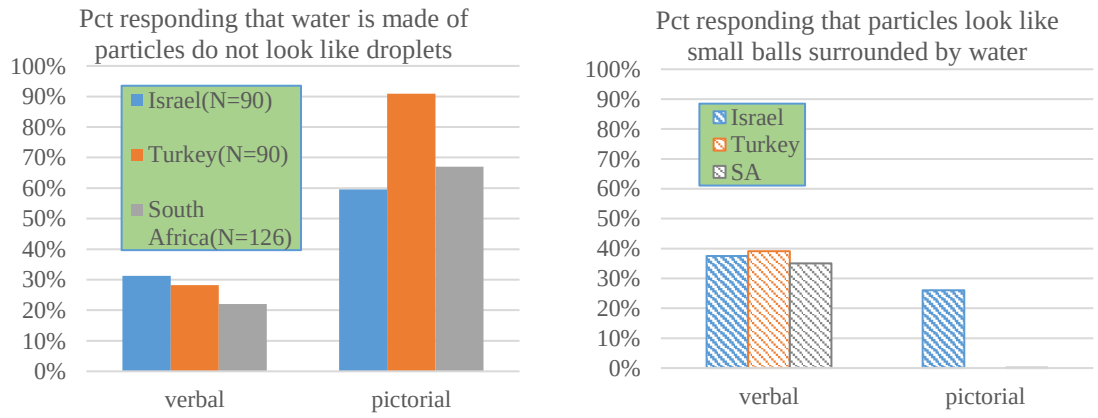


Figure 4. Response rates to item 1 - Water Droplet. Note the higher proportion of correct responses in the pictorial format in all three groups (left), and the lower percentage of the most common inappropriate idea (B) in which water is represents balls immersed in a liquid in the pictorial format (right).

Figure 4 shows that a majority of students from all three groups chose response "E" representing the normative scientific model of water in the pictorial form, but a much smaller proportion chose the equivalent verbal response as shown in Figure 4. Conversely, Figure 6 shows that in item 8, addressing the apparent "disappearance" of sugar upon dissolving in water - the appropriate pictorial option was chosen by a significantly smaller proportion of students than the verbal one. Figure 5 shows the proportion of students who chose the correct description of the configuration of the remaining air particles in the balloon. This proportion was similar in the pictorial and verbal formats, except in the Israeli group.

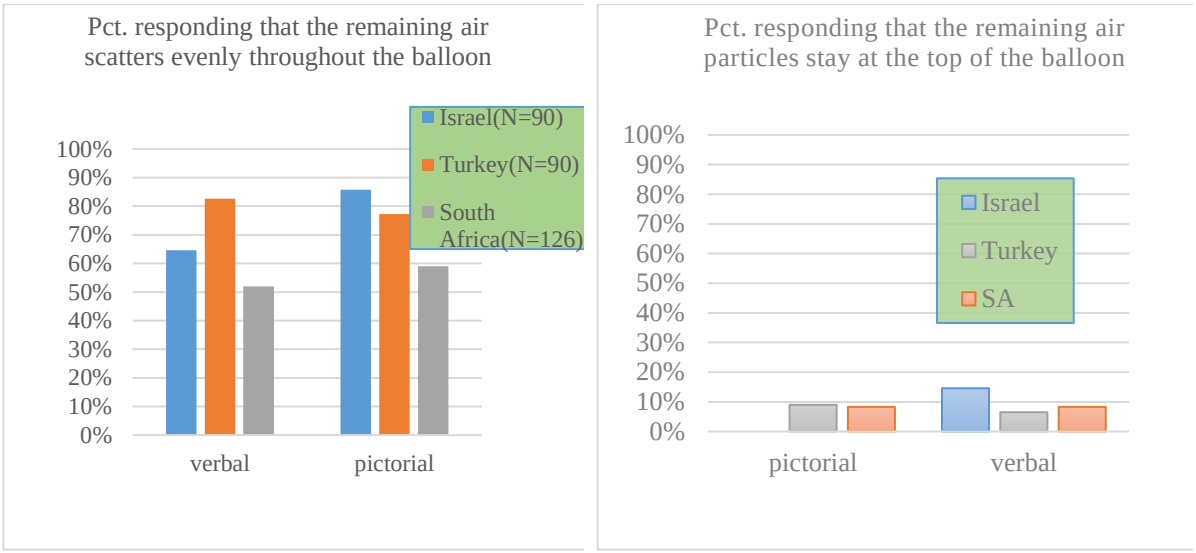


Figure 5. Response rates for item 5 – arrangement of remaining air particles in a balloon. In this item, there were no significant differences in the proportion of students who chose the correct response in the verbal and pictorial formats, except in the Israeli sample, which performed significantly better in the pictorial format, and did not choose the "stay at the top" option in the pictorial format.

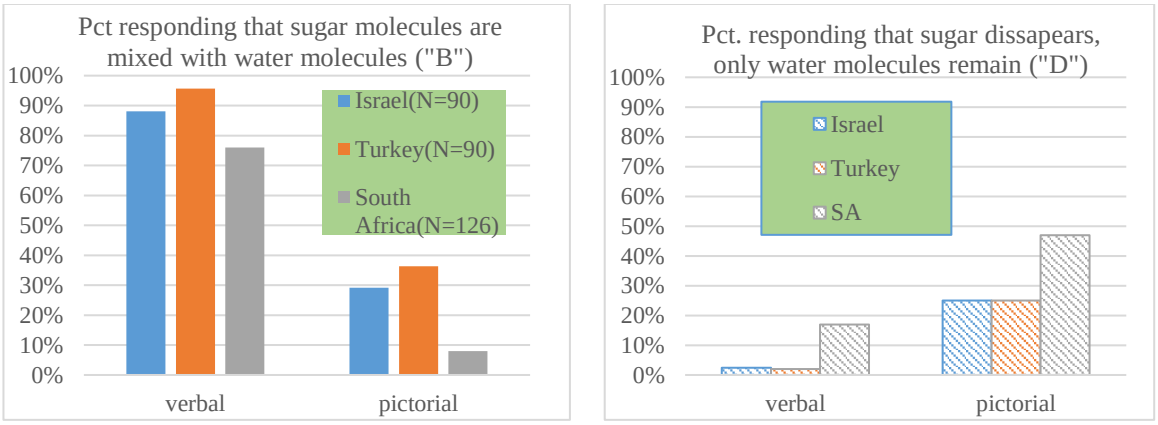


Figure 6. Response rates for item 8 – sugar in water. Note the higher proportion of students who chose the correct verbal response than those who chose the pictorial one (left). Many of those who did not choose the correct picture, chose option "D" in which only water molecules are present (right). Fewer students chose the equivalent verbal response.

The proportions of correct responses in each format of the survey are summarized in Table 1 with a chi-square analysis of the significance of the differences in proportions between formats. Note the significantly higher proportion of students who chose the correct pictorial response in item 1 and the correct verbal response in item 8. The differences in item 5 are not significant, except in the Israeli group.

Table 1. Within country differences between verbal and pictorial formats.

ITEM	*COUNTRY*	*VERBAL FORMAT (PCT. CORRECT)*	*PICTORIAL FORMAT (PCT. CORRECT)*	χ^2*(SIG.)*
1	Israel (N=90)	31.3%	59.5%	7.25 (0.008)**
	Turkey (N=90)	28.3%	90.9%	35.45(<0.001)***
	South Africa (N=126)	21.7%	66.7%	25.7(<0.001)***
5	Israel (N=90)	64.6%	85.7%	5.76 (0.024)**
	Turkey (N=90)	82.6%	77.3%	0.40 (0.52)
	South Africa (N=126)	52.0%	59.1%	0.70 (0.40)
8	Israel (N=90)	88.1%	29.2%	31.7(<0.001)***
	Turkey (N=90)	95.7%	36.4%	50.44(<0.001)***
	South Africa (N=126)	75.8%	8.3%	75.6(<0.001)***

P<0.05 ** P<0.01***

Between-country differences

In addition to comparing the verbal and pictorial formats, we examined also the between-country differences within each format (Table 2). Interestingly, we found significant differences between the countries within the pictorial format questions: In item 1 – the Turkish group performed significantly higher than the Israeli and South African samples, and in items 5 and 8, the South African group performed significantly lower than the Israeli and Turkish groups. In the verbal format, the only significant difference was found in item 5, in which the South African group performed significantly lower than the Israeli and Turkish groups.

Table 2. Between countries differences within verbal or pictorial formats.

Item	*Verbal Format* $\boldsymbol{\chi^2}$*(sig.)*	*Pictorial Format* $\boldsymbol{\chi^2}$*(sig)*
1	1.34 (0.51)	11.8(0.002)***
5	7.1(0.029)**	9.92(0.007)***
8	5.58 (0.07)	9.55 (0.008)***

Analysis of the second tier

Students were asked to explain their pictorial choice verbally, or vice-versa. For example, a student chose the pictorial option D for item 8 (dissolving sugar) and explained: “The sugar cube dissolved in the hot water to form a solution, which is why it is not seen”. This student seemed to understand the term “dissolving” as “disappearing in water”, which reflects her observation at the macro level. Note that the scientifically appropriate verbal response to this question “The sugar dissolves, and the sugar particles mix with the water particles” is the only

one that contains the term "dissolving". It might be that many students chose this response because this word served as a verbal cue which signified a scientific term they heard in class. These students were not necessarily familiar with the underlying molecular level model of the mixture.

The drawings made by the students in response to item 1 reveal that many of them were familiar with the molecular representation of water molecules. Figure 7 (left) shows a drawing by South African student who chose option E ("There are many tiny water particles in a drop of water that do not look like droplets"). Conversely, the drawing in Figure 7 (right) represents an incomplete model of a student who chose the correct verbal option B. In this drawing of the dissolved sugar, one of the components –either water molecules or sugar molecules – is missing. This again illustrates that students who chose the correct verbal response, used the verbal cue in the question although their molecular-level model of the structure of a mixture – was lacking.

Figure 7. Students' drawings that explain their choices to item 1 (left), and to item 8 (right)

DISSCUSION

We developed two formats of the same conceptual questionnaire about the particle model of matter. We randomly assigned questionnaire format to students in three countries and found significantly different response patterns in the verbal and pictorial format of half of the questionnaire items. Thus, what seems to be equivalent verbal and pictorial representations, failed to capture the same "reasoning levels" or mental models among students. This phenomenon is consistent across three groups of students from three countries, although some differences occur due to differences in curriculum and student populations.

On average, the between-country differences in the verbal format were smaller than the pictorial format. This might indicate that the questions in the pictorial format were less reliable. However, the significant differences between the South African group and the other two groups in interpreting the pictorial format might stem from the difference in curricular activities. For example, only 13% of the South African students reported that their teachers used particle simulations, whereas in Israel and Turkey the vast majority of the students reported using simulations (96% and 88% respectively).

In order to examine the origin of the difference between formats, we triangulated the findings from the multiple choice questions with the students' own drawings and written explanations. We suggest that differences between the pictorial and verbal formats stem from molecular-level cues that were apparent in the pictorial format and missing from the verbal one or vice-versa. The pictorial format in item 1 (see Figure 1) elicited a familiar representation - the molecular structure of water - while the verbal response did not. The structure of the water

molecule was familiar to many students – especially in Turkey – where the molecular structure is part of the curriculum already in 7th grade. Conversely, the pictorial format in item 8 (see Figure 3) contained unfamiliar information (the hexagonal sugar molecule in option B) that deterred students from choosing this option – and led many of them to choose option D.

CONCLUSION

Our study shows that descriptions of students' levels of reasoning based on their responses to multiple-choice items in studies such as Hadenfeldt et al., (2016) should be taken with caution. We suggest that in order to make more substantiated inferences, studies should rely more on sets of two or more items that examine the same concepts using verbal and pictorial information. Only respondents who use the same reasoning level in the pictorial item and the verbal item, can be considered as using a coherent and stable "mental model".

REFERENCES

Akaygun, S. & Jones, L. L. (2014). Words or Pictures: A comparison of written and pictorial explanations of physical and chemical equilibrium. *International Journal of Science Education*, *36*(5), 783-807

Hadenfeldt, J. C., Neuman, K., Bernholt, S., Liu, X., & Parchman, I. (2016). Students' progression in understanding the matter concept. *Journal of Research in Science Teaching*, *53*(5), 667–708.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. P. Mestre (Ed) Transfer of Learning from a Modern Multidisciplinary Perspective. Greenwich, CT: IAP. (pp. 89-119)

Johnson, P. M. (1998) Progression in children's understanding of a 'basic' particle theory: a longitudinal study. *International Journal of Science Education*, *20*, 393—412

Langbeheim, E. (2015). Reinterpretation of students' ideas when reasoning about particle model illustrations. *Chemistry Education Research and Practice*, *16*(3), 697-700.

Merritt, J., & Krajcik, J. (2013). Learning progression developed to support students in building a particle model of matter. In *Concepts of matter in science education* (pp. 11-45). Springer Netherlands

Nussbaum, J., & Novick, S. (1982). Alternative frameworks, conceptual *conflict and acco*mmodation: Toward a principled teaching strategy. *Instructional science*, *11*(3), 183-200.

Nyachwaya, J. M., Mohamed, A.R., Roehrig, G. H., Wood, N. B., Kern, A. L., & Schneider, J. L. (2011). The development of an open-ended drawing tool: An alternative diagnostic tool for assessing students' understanding of the particulate nature of matter. *Chemistry Education Research and Practice*, *12*, 121–132.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(2), 115-163.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

DEVELOPMENT OF A TOOL TO ASSESS SECONDARY SCHOOL STUDENTS' UNDERSTANDING OF MEASUREMENT UNCERTAINTIES

Johannes Schulz[1], *Burkhard Priemer*[1] *and Amy Masnick*[2]
[1] Humboldt-Universtät zu Berlin, Berlin, Germany
[2] Hofstra University, Hempstead NY, U.S.A.

Estimating the quality of evidence is a core competence in science and science education. Applying this to quantitative observations means, for example, to judge the uncertainties in measurements. In order to do this systematically, a reference frame is needed. Hellwig (2012) and Priemer and Hellwig (2016) present a model that describes and structures content about measurement uncertainties relevant for the secondary school level. With a reference to this model, we are developing an assessment tool to probe students' understanding of measurement uncertainties. This is done by formulating competencies and interpreting these as latent constructs. For these constructs, scales were developed based on Item Response Theory. This paper describes the general approach to the development of the tool and illustrates it with the two example scales: "Reliability of a Measurement Result" and "Comparison of a Result with other Values".

Keywords: measurement, assessment of competence, secondary Education

INTRODUCTION

Competencies in identifying and handling measurement uncertainties are necessary to understand and perform scientific work like analyzing empirical data from experiments. The related activities can be identified in models that describe inquiry processes (e.g. the phase of data interpretation; Pedaste, Mäeots, Siiman, De Jong, Van Riesen, Kamp et al., 2015), or experimental work (e.g. "evidence evaluation" in the Scientific Discovery as Dual Search (SDDS) model; Klahr & Dunbar, 1988). If key concepts about judging the quality of data – or more specifically estimating measurement uncertainties – are important for teaching, we also need assessment tools that probe students' competencies accordingly. This paper describes the development of such a tool. Since the development of the tool is not yet complete, we focus on example subscales. These subscales are related to a comprehensive framework and illustrate the characteristics of the test.

THEORETICAL BACKGROUND

Experimental work with measurements is seen as an important practice in science education and, hence, appears in national science curricula (in the USA: NGSS Lead States, 2013, Next Generation Science Standards [NGSS], Appendix F, Practice 4; in England: Department for Education and Employment and Qualifications and Curriculum Authority, 19 99, Science The National Curriculum for England, p. 37-38; in Germany: KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK], 2004, Bildungsstandards im Fach Physik für den Mittleren Schulabschluss, p. 11). Whenever there is measurement, there can be variability in the measurement, and thus, uncertainty about how to assess the resulting data. To judge the quality of collected data, it is often essential to capture

and discuss measurement uncertainties, and it is difficult to know how to address this issue without training. However, measurement uncertainties are an oft-neglected topic in science education (Hellwig, 2012). One concern is that teachers do not have guidance on the most important issues in measurement evaluation and the best ways to teach such evaluation (compare i.e. Priemer & Hellwig 2016). Hence, research is needed that develops and analyzes learning progressions and teaching instructions to bring this topic to teachers' attention and to facilitate effective teaching of these concepts.

To learn about students' understanding of measurement uncertainties, it's important to have a way to assess their knowledge. Although there is some research on the development of teaching instructions (as outlined for example in Deardorff, 2001, and Munier, Merle, & Brehelin, 2011) and investigations about students' views in this field (as in Buffler, Allie, Lubben, & Campbell, 2001, and Masnick & Morris, 2008), validated instruments are needed that assess students' understanding about measurement uncertainties. Additional assessments of understanding about nature of measurements have been developed by Day and Bonn (2011), Lubben and Millar (1996), Garratt, Horn, and Tomlinson (2000), and Volkwyn (2005). However, most of these existing instruments address upper secondary or university education and focus on certain subtopics of measurement uncertainties, e. g. the reliability of data (Lubben & Millar, 1996). Hence, a tool that investigates secondary school students' comprehensive understanding of concepts of measurement uncertainties is still missing. It is the goal of this paper to outline the development of such a tool. We do so by describing the general stages of tool development, and then detailing two of the instrument's concepts and how they were assessed.

The development of this tool is based on a framework from Hellwig (2012) and Priemer and Hellwig (2016), who presented a comprehensive content structure model for the field of measurement uncertainty for secondary and for university education. The content is structured in four main dimensions and ten concepts for both educational levels (Table 1). Hellwig (2012) also developed subconcepts and subsubconcepts (overall more than 50) for each concept, which are not described here to keep the paper readable.

RESEARCH QUESTIONS

To develop a tool to assess secondary school students' understanding of measurement uncertainties, we posed the following questions:

1. How can the concepts of the framework model be operationalized and measured?
2. What is the quality of the developed scales?

In this paper, we answer these questions for the concepts "Reliability of a Measurement and the Result" and "Comparison of a Result with other Values".

Table 1. A content structure model for the field of measurement uncertainties (Hellwig, 2012; Priemer & Hellwig, 2016).

Dimension	Concept
Existence of Uncertainties	Sources of Uncertainty
	Distinguishing Uncertainty from Error
Handling of Uncertainties	Measuring Objective
	Result of a Measurement
Assessment of Uncertainties	Direct Measurement: Evaluating a Single Uncertainty Component
	Indirect Measurement: Propagation of Uncertainty
	Expanded Uncertainty
Conclusiveness of Uncertainties	Reliability of a Measurement and the Result
	Comparison of a Result with other Values
	Fitting Data to a Straight Line / Fitting Data to an Expected Curve

METHOD

The development of the tool followed four steps: 1. the formulation of competencies for all concepts of the framework model (Table 1); 2. the operationalization of these competencies in test items; 3. the assessment of the validity of the items; and 4. an empirical test of the scales that represent the concepts.

Step 1: Formulating competencies for the concepts

For each of the ten concepts of the framework model (Table 1), competencies were formulated based on the content suggested in Hellwig (2012) with additional consideration of the *Guide to the expression of uncertainty in measurement* (GUM; Joint Committee for Guides in Metrology, 2008). The competencies were developed by analyzing the content of the concepts and describing performances expected by students who are familiar with the corresponding content. This was done by an expert in the field of science education, and the competencies and their assignment to concepts were validated by another expert. In this step, we also made sure that each concept consists of unique content and that there are no overlaps in the content of different concepts. Table 2 lists the competencies for the two concepts “Reliability of a Measurement and the Result” and “Comparison of a Result with other Values”.

Step 2: Operationalizing the competencies in test items

The competencies were used to develop test items by choosing specific situations, experiments, and tasks that are relevant for secondary school instruction. For each concept, we added an introduction page to the test booklet which gives an overview of the content to assure that students understand the scientific terms used. For example, one introduction page explained how the overall uncertainty of a measurement is estimated when a number of different uncertainties are given that all influence the measurement (the uncertainty budget). Further,

most of the concepts are illustrated with additional examples. This information was given to make sure that the test assesses students' competencies *using* the concepts instead of simply remembering facts.

Table 2. Competencies for two concepts of the framework model.

Concept	Competencies
Reliability of a Measurement and the Result	The students are able to…
	… present a result of a measurement with measurement uncertainties (using the correct numbers of decimals)
	… give an uncertainty budget and interpret it with regard to the size of the different uncertainty effects
	… judge the reliability of a measurement based on the uncertainty budget and the correct number of decimals
Comparison of a Result with other Values	The students are able to…
	… compare the result of measurement with a reference value by analyzing its position with respect to the interval of the uncertainty
	… compare two or more measurement results by analyzing the intersections of their intervals of uncertainty
	… identify outliers in measurements and discuss them according to their influence on the result of a measurement
	… compare the intervals of the uncertainty for different sample sizes

All of the items were designed in multiple choice format (with only one correct answer) or multiple answer format (where more than one answer could be selected, and there was at least one wrong answer and at least one correct answer). Figure 1 gives an example of a test item that addresses the competence "The students are able to judge the reliability of a measurement based on the uncertainty budget and the correct number of decimals" of the concept "Reliability of a Measurement and the Result". We created 17 items each for the two concepts "Reliability of a Measurement and the Result" and "Comparison of a Result with other Values". We will focus on them in this paper. The full instrument includes 150 items assessing all 10 concepts.

Step 3: Assessing the validity of the test items

In order to validate the items of the complete model with the ten concepts we created an item-subset including item designed to assess each of the ten concepts. We presented this subset of 52 items to three experts in the field of metrology together with a list of all competencies for all concepts, and asked them to assign the items to the concepts. The items that were given to the experts were chosen at random with the exception that there was at least one item in the subset for every concept. The restriction to a subset of items was necessary due to time limitations. For the two concepts mentioned above, six items were included. We also added an eleventh category to the expert rating for items that did not fit into any of the provided categories. Finally, the experts had room to give comments.

Scales

In an experiment, Brian wants to measure the mass of an object as precisely as possible. He uses a calibrated scale (left) and an uncalibrated kitchen scale (right). That means that the kitchen scale was not tested with respect to its accuracy by comparing it with a standard scale. "Zero position" indicates how accurately the scales read 0 g when nothing was put on it. For both scales he identified possible sources of uncertainty and listed these in an uncertainty budget. The overall uncertainty was determined by rounding.

calibrated scale

kitchen scale

Source of Uncertainty	Uncertainty budget
Display digits	0.001 g
Calibration	0.1 g
Zero position	0.1 g
Overall uncertainty	0.2 g

Source of Uncertainty	Uncertainty budget
Display digits	0.01 g
Calibration	Not calibrated, uncertainty unknown
Zero position	0.1 g
Overall uncertainty	0.1 g

Which of the two scales is more reliable and why?

□ The calibrated scale is more reliable because the precision of the display digits is much better.

□ Both scales are equally reliable because their uncertainty of the zero position is the same.

□ The kitchen scale is more reliable because it has the smaller overall uncertainty.

x It is not possible to compare the reliabilities of the two scales because the uncertainty of the calibration of the kitchen scale is unknown.

Figure 1. Example item "scales" of the concept "reliability of a measurement and the result". The correct solution is marked: without calibrating the uncertainty budget is incomplete and hence the two scales cannot be compared.

Step 4: Empirical test of the scales that represent the concepts

We presented the 34 items of the two concepts "Reliability of a Measurement and the Results" (17 items) and "Comparison of a Result with other Values" (17 items) to 143 pupils from the 8th grade to the 12th grade in six different classes of three German schools in an urban area. All students were asked to answer all items. The order of the items was the same within the two concepts, but half the participants saw questions about one concept first, and half saw the other concept first. The participants had as much time as they needed to answer. No student took longer than 90 minutes.

RESULTS

Results of Step 3: Assessing the validity of the test items

In the first empirical step of item development, the three experts gave the following rating to the 52 items: 31 items were sorted to the same concept by all three experts, 14 items were sorted to the same concept by two of the three experts, and 7 items were sorted to three different concepts by the three experts. The inter-rater agreement was $\kappa = 0,67$ (Fleiss' kappa). When restricting the inter-rater agreement to the two concepts in focus (with the six items) we obtained $\kappa = 0,50$ (three items were assigned to the same category by all three experts, two items were assigned to the same category by two experts, and one item was assigned to three different categories by the experts). All items were kept. However, items that fell into two or more categories were modified based on the experts' comments and based on a discussion between the authors.

Results of Step 4: Empirical test of the scales that represent the concepts

Next, we analyzed the pupils' responses to the 34 scale items. A Rasch analysis and additional tests were calculated with Winsteps and R (also R Studio). We chose Item Response Theory (IRT) instead of Classical Test Theory (CTT) because IRT has stricter conditions regarding the characteristics of the items and because it allows us to display the estimated ability of the students and the difficulty of the items on the same scale (for an introduction to IRT see for example Hambleton, Swaminathan, & Rogers, 1991). Figure 2 shows the Wright maps for the concepts "Reliability of a Measurement and the Results" and "Comparison of a Result with other Values". The left side of each diagram displays the number of participants that reached a certain ability in a frequency diagram. This ability was scaled by algorithm of the program R (in this case here from -3 to 4). This scale is also used to assign each of the items a certain difficulty. The right side of each diagram shows these difficulties in ascending order. Thus, we can display the estimated ability of the students and the difficulty of the items in one single diagram.

The difficulties of the items lie between -0,93 and 2,70 (Reliability of a Measurement and the Results) and between -1,08 and 1,26 (Comparison of a Result with other Values). The estimated abilities lie between -2,41 and 1,99 and -2,74 and 4,45 for these concepts, respectively. The Expected a Posteriori (EAP) reliability of the Rasch-analysis for the concept "Reliability of a Measurement and the Results" is $r = 0,54$ and for the concept "Comparison of a Result with other Values" $r = 0,80$. For the WLE reliability we computed 0,75 (Comparison of a Result with other Values) and 0,49 (Reliability of a Measurement and the Results).

We also looked at the Unweighted Mean Square (MNSQ) Outfit values, which indicate how accurately or predictably the data fits the Rasch model. We decided to use the MNSQ-Outfit over the MNSQ-Infit since the MNSQ Outfits are more sensitive to items with difficulty far from the estimated ability of the participant. The MNSQ-Outfits are in the range between 0,92 and 1,18 (see Figure 3) for the concept "Reliability of a Measurement and the Results" and between 0,76 and 1,37 for the concept "Comparison of a Result with other Values" (Figure 3). Thus, the MNSQ Outfit values of all but one item are inside the interval of 0,7 - 1,3

recommended by Linacre and Wright (1994) and are therefore useable for a measurement (see also Linacre & Wright, 1994).

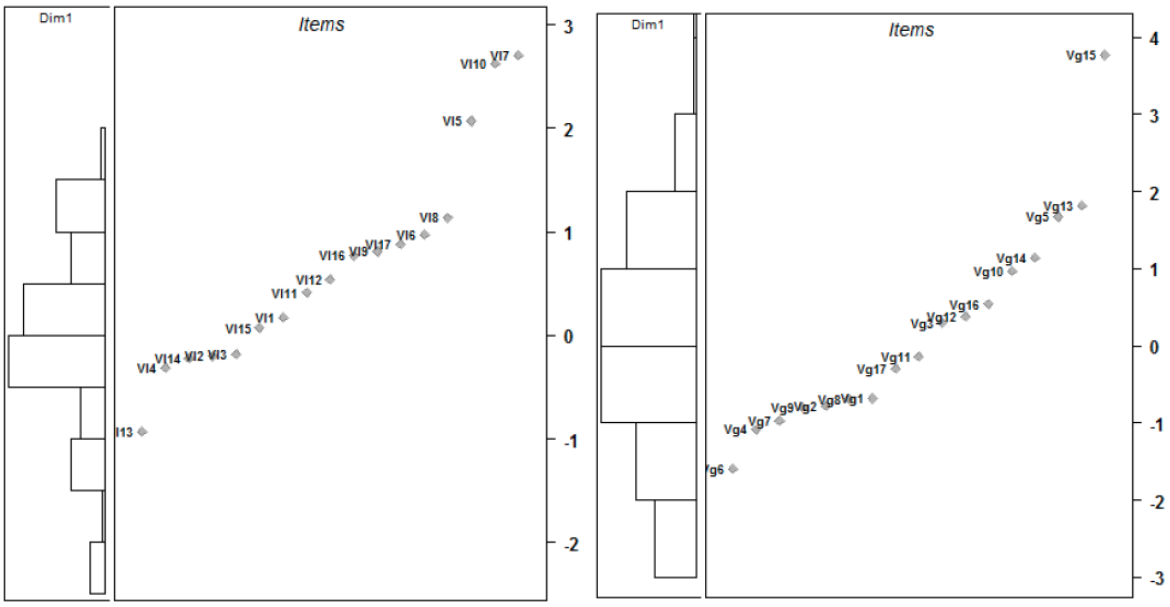


Figure 2. Wright-Maps for the concepts "Reliability of a Measurement and the Results" (left) and "Comparison of a Result with other Values" (right); all 17 items of both concepts were answered by n=143 students.

Item	estimate	MNSQ Outfit	Item	estimate	MNSQ Outfit
Vl1	0,17	0,99	Vl10	2,62	1,18
Vl2	-0,20	0,99	Vl11	0,41	1,06
Vl3	-0,18	0,98	Vl12	0,54	1,01
Vl4	-0,31	1,03	Vl13	-0,93	0,93
Vl5	2,07	1,11	Vl14	-0,22	1,03
Vl6	0,97	0,92	Vl15	0,07	0,93
Vl7	2,70	0,93	Vl16	0,77	1,01
Vl8	1,13	0,96	Vl17	0,88	0,98
Vl9	0,81	1,04			

Item	estimate	MNSQ Outfit	Item	estimate	MNSQ Outfit
Vg1	-0,68	1,37	Vg10	0,95	0,95
Vg2	-0,78	1,02	Vg11	,095	0,95
Vg3	0,30	0,76	Vg12	0,94	0,94
Vg4	-1,08	0,78	Vg13	0,76	0,76
Vg5	0,85	0,85	Vg14	0,84	0,84
Vg6	0,79	0,79	Vg15	1,09	1,09
Vg7	0,99	0,99	Vg16	0,98	0,98
Vg8	0,81	0,81	Vg17	1,18	1,18
Vg9	1,26	1,26			

Figure 3. Estimated Difficulties and MNSQ Outfits for the items of the concepts "Reliability of a Measurement and the Results" (top) and "Comparison of a Result with other Values" (bottom).

DISCUSSION

Assessing the validity of the test items

From the metrology experts' ratings, we evaluated how well each item appeared to measure the concept it was intended to measure. According to Landis and Koch (1977), the inter-rater agreement of the experts can be interpreted as "substantial" (for all 52 items) and "moderate" (for the six items of the two concepts in focus). This result has to be seen in light of the restriction that the experts had eleven categories to choose from, that only a subset of the developed items of the complete test was rated, and that only six items of the two concepts in focus were included. However, the expert rating indicated that although there was agreement on many items, some of the items needed further improvement. One expert suggested to be clearer in the use of technical terms. For example, the item shown in Figure1 must differ between gauging and calibrating. Further, some of the experts remarked that some of the answer options of specific items fell into different categories. This problem was addressed by choosing more fitting answer options for those items. To verify that all the changes of the items are improvements, and lead to clear mappings between items and concepts, a second round of expert rating is needed.

Concerning the method of the expert rating, we have to keep in mind, that assigning the items to the concepts by experts is only one way of estimating the validity. This validation procedure is also limited in the way that it can't ensure that the content of the concept is covered completely by the items.

Empirical test of the scales that represent the concepts

The items generally fit the Rasch model. That means that the strict conditions that the IRT specifies for the items is fulfilled. The MNSQ Outfit values of all but one item are inside the interval of 0,7 - 1,3 recommended by Linacre and Wright (1994). The MNSQ outfit value is marginally beyond the threshold (with 1,37 as shown in Figure 3) for one item (Vg 1) only. That means from a statistical point of view that the item may be unproductive for the scale but it is not degrading the measurement system. Since the deviation is very small and since this item received high agreement by all three experts, we kept it in the test. The Wright maps for the two concepts show that the distribution of the difficulty of the items fit the competencies of the students quite well. However, items with lower difficulty can improve the scale "Reliability of a Measurement and the Result". Those items could replace some of the many items with a medium difficulty, so the test could cover a greater spectrum of difficulty without increasing the number of items.

For the difference in the reliability of both concepts there might be several reasons. Most obvious it might be possible that the items of the concept "Comparison of a Result with other Values" need further improvement. But if we keep in mind that the concepts are derived from a model which includes subconcepts for each concept, it might also be possible that the concepts split up in two (or even more) subdimensions. For example it might be possible that for the participants, the presentation of a measurement result with uncertainty is another competency, different from giving and interpreting an uncertainty budget (see competencies Table 2). Further research is needed to come to a final judgement here. Currently we are

working on further analysis of the data, for example a detailed Rasch analysis including an analysis of potential subdimensions.

To assure that including new items reduces the gap in the Wright map, it is necessary to test the set of items again. This should be useful anyway since experts recommended changes to some of the items. If those changes also affect the reliability of the items, they must also be controlled when testing further improved items in a future study.

CONCLUSIONS

The results show that our test instrument to assess secondary school students' understanding of measurement uncertainties works well for the concepts discussed. The scales and items have desired levels of difficulty and cover students' competencies (with a few exceptions). Even though there is room for improvement (e.g. in the validity and the range of the difficulty of the items), the results of this study show that the concepts of the model can be measured well using multiple choice items. These findings show that the process of developing items for each concept, having the items evaluated by metrology experts, and then tested for coherence by actual students, is a productive method of developing and validating a scale for assessing understanding measurement uncertainty. We are currently working on improvements of the items and on the development of scales for all ten concepts.

ACKNOWLEDGEMENT

We would like to thank Sarah Heydemann and Laura Kemnitzer for their support in the study.

REFERENCES

Buffler, A., Allie, S., Lubben, F., & Campbell, B. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education, 23*(11), 1137-1156.

Day, J., & Bonn, D. (2011). Development of the concise data processing assessment. *Physical Review Special Topics - Physics Education Research, 7*, 010114. Doi: 10.1103/PhysRevSTPER.7.010114

Deardorff, D. (2001). Introductory physics students' treatment of measurement uncertainty (Diss., North State University, Raleigh, NC). https://www.ncsu.edu/PER/Articles/DeardorffDissertation.pdf

Department for Education and Employment & Qualifications and Curriculum Authority (1999). Science The National Curriculum for England. Retrieved from http://dera.ioe.ac.uk/4402/1/cSci.pdf

Garratt, J., Horn, A., & Tomlinson, J. (2000). Misconceptions about error. *University Chemistry Education, 4*(2), 54-57.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). Fundamentals of Item Response Theory. London: Sage.

Hellwig, J. (2012). Messunsicherheiten verstehen – Entwicklung eines normativen Sachstrukturmodells am Beispiel des Unterrichtsfaches Physik. Dissertation: http://www-brs.ub.ruhr-uni-bochum.de/netahtml/HSS/Diss/HellwigJulia

Joint Committee for Guides in Metrology (2008). Evaluation of measurement data - Guide to the Expression of Uncertainty in Measurement (GUM). Paris, France: Sèvres Cedex.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48. doi:10.1207/s15516709cog1201_1

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). Bildungsstandards im Fach Physik für den Mittleren Schulabschluss

[Science standards for middle school graduation for the school subject physics]. München: Wolters Kluwer.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data, *Biometrics, 33, 159-174*.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values, http://www.rasch.org/rmt/rmt83b.htm

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, *18*(8), 955-968.

Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79*, 1032-1048. doi:10.1111/j.1467-8624.2008.01174.x.

Munier, V., Merle, H., & Brehelin, D. (2011). Teaching scientific measurement and uncertainty in elementary school. *International Journal of Science Education*, iFirst Article, 1-32, doi: 10.1080/09500693.2011.640360

NGSS Lead States (2013). Next generation science standards: For states, by states. Washington, DC: The National Academies Press.

Pedaste, Mäeots, Siiman, De Jong, Van Riesen, Kamp et al. (2015). Phases of inquiry-based learning: definitions and the inquiry cycle. *Educational Research Review, 14*, 47-61, https://doi.org/10.1016/j.edurev.2015.02.003

Priemer, B., & Hellwig, J. (2016). Learning about measurement Uncertainties in secondary education: A model of the subject matter. *International Journal of Science and Mathematics Education, 16*(1), 45-68, doi:10.1007/s10763-016-9768-0.

Volkwyn, T. S. (2005), First year students' understanding of measurement in physics laboratory work. Dissertation at University of Cape Town.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

THE DESIGN AND IMPLEMENTATION OF AN ASSESSMENT METHOD COMBINING FORMATIVE AND SUMMATIVE USE OF ASSESSMENT

Sanne Schnell Nielsen[1], Jens Dolin[1], Jesper Bruun[1] and Sofie Birch Jensen[2]
[1]University of Copenhagen, Copenhagen, Denmark
[2]King's College, London, England

The two key purposes of assessment, formative and summative, often contradict each other when attempted used simultaneously. Summative assessment of learning will generally prevent formative assessment for learning. Moreover, in order to be useful and manageable, the assessment method must be easy to integrate into and should ordinary teaching. This study explores how such an assessment method, called the Structured Assessment Dialogue (SAD), could be designed and the rationale behind it. Then the study investigates the challenges and benefits perceived by teachers related to the uptake of SAD in their daily practice. The SADs were undertaken in science, technology and mathematics in primary and secondary schools in Denmark and Finland. The data used in this study include teacher-generated preparation and reflections forms, interviews with teachers, and an open-ended questionnaire. Our findings suggest that SAD holds prospects for fulfilling the purposes for both formative and summative assessment, with the highest prospects related to formative assessment purposes and characteristics. However, it needs time, change of classroom culture, and adjustments and careful implementation and routine building. In addition, teachers must be proficient in addressing the different aspects and levels of competences throughout the SAD.

Keywords: assessment methods, assessment of competence, inquiry-based teaching

INTRODUCTION

The two key purposes of assessment, formative and summative, often contradict each other when attempted used simultaneously. Summative assessment of learning will generally prevent formative assessment for learning to be realised, so the learning potential of the assessment will often be minimal (Butler, 1988). It is therefore interesting to find ways to combine the dual use of assessment, which do not diminish the learning potential. Moreover, in order to be useful and manageable, the assessment method must be easy to integrate into and should align with ordinary teaching.

Such an assessment method, called the Structured Assessment Dialogue (SAD), has been developed as part of a European research project Assess Inquiry in Science, Technology and Mathematics Education (ASSIST-ME). SADs are ritualized, short, 3-part assessment activities integrated as part of ordinary classroom teaching. SADs aim at uncovering student competencies while also providing students with feedback and time for self-reflection.

The next section describes different purposes of formative and summative assessment and elaborates on the characteristics they have in common as well as their differences.

Purposes and characteristics of formative and summative assessment

The differences between formative and summative use of assessment are pinpointed in wordings: Assessment *for* learning and assessment *of* learning. The same assessment method can be used for both formative and summative purposes but the use defines whether the assessment is formative or summative. Formative use of an assessment method is meant to improve students' learning (or teachers' teaching). Summative use of the same assessment method will judge students' level of competence (or teachers' teaching). We see summative and formative assessment as inherently linked, but with formative assessment focusing on student involvement, student achievement of both normative and personal criteria, and finding the next learning step (Black and Wiliam, 1998; Harlen, 2012, 2013; Dolin, Black, Harlen and Tiberghien, 2018a). Both uses of assessment rely on evidence of student performance but while the interpretation of this evidence is criterion-referenced (i.e. related to the learning goals) for summative purposes, it is also student-referenced for formative purposes. This is because students need to know what to learn next and how to do it. This is student-specific and needs student involvement to be realised. The other critical aspect of formative assessment is 'finding the next learning step' in a series of learning steps; a progression.

Just like summative and formative assessment can be seen as part of the same cycle, they can also be seen as two ends of the same continuum, rather than a dichotomy. This illustrated in Figure 1.

Formative<----------->Summative				
	Informal formative	**Formal formative**	**Informal summative**	**Formal summative**
Major focus	What are the next steps in learning?		What has been achieved to date?	
Purpose	To inform next steps in learning	To inform next steps in learning and teaching	To monitor progress against plans	To record achievements of individuals
How is evidence collected	As normal part of class work	Introduced into normal class work	Introduced as a special part of normal class work	Separate task or test
Basis of judgement	Student- and criterion-referenced	Student and criterion-referenced	Student and criterion-referenced	Criterion-referenced

Figure 1. Formative and summative assessment as a continuum. The SAD is placed as the rectangle overlapping both.

The placement of the SAD as being able to have elements of both formative and summative assessment will be substantiated in the following.

Rationale for designing a classroom and dialogue-based assessment method

Our design of a dialogue-based assessment draws upon the Norwegian researcher Olga Dysthe (1996), seeing dialogue as a central way to learning. Dysthe is inspired by the Russian linguist Bakhtin (1981), and the key point is to open a room for student reflection in a non-authoritative environment.

Teacher-led classroom dialogue is one of the most common instruction practices worldwide (Wiliam & Leahy, 2015). Moreover, a large proportion of the information that teachers obtain through informal formative assessment is obtained through classroom dialogue (Ruiz-Primo, 2011). Hence, introducing a dialogue-based assessment method in the classroom will to a large extent align with an already existing instruction and assessment approach.

When we designed the SAD, the aim was two pronged: on the one hand we wanted to develop a method resembling an already existing dialogue-based formative assessment practice. On the other hand, we also wanted to develop a method that teachers could use for both formative and summative assessment.

Most formative assessments within the typical classroom are quite informal in nature, generally unplanned and used differently by different teachers (Shinn, 2013). However, in order for an assessment method to be able to contribute to summative purposes it must provide a relatively standardized approach to how it is administered.

To address the above factors, the format of the SAD was intended to be structured, planned and formal. This is reflected in the SAD design by establishing a concise phase and time structure, well defined assessment criteria, and an unequivocal division of roles among the participants in the classroom. In the next section, we will briefly describe the structure of the SAD. A more detailed description of SAD and its operationalization can be found in Dolin, Bruun, Nielsen, Jensen and Nieminen (2018b).

In order to be effective, formative assessment has to be integrated into classroom practice (Wiliam, 2011). The SAD is intended to be an integrated component of the ordinary teaching, yet the setting is different from the ordinary teaching. As part of the setting, each student gets appointed to undertake a specific role (i.e. focus student, feedback student or self-reflecting student) and, subsequently, all students are physically rearranged according to these roles (Figure 2).

The SAD consists of three distinct phases, each with a clear and well-defined role for the teacher and for each of the students. The three phases have specific functions related to formative and summative assessment:

Phase 1. An assessment dialogue between the teacher and a preselected focus student based on specific assessment criteria related to a specific aspect of a competence (5-7 minutes).
Phase 2. Peer feedback from a group of students to the focus student (5 minutes).
Phase 3. Self-assessment where all students in the classroom assess their own level and reflect on their learning needs (2-3 minutes). For summative purposes, the teacher notes down the level of the focus student.

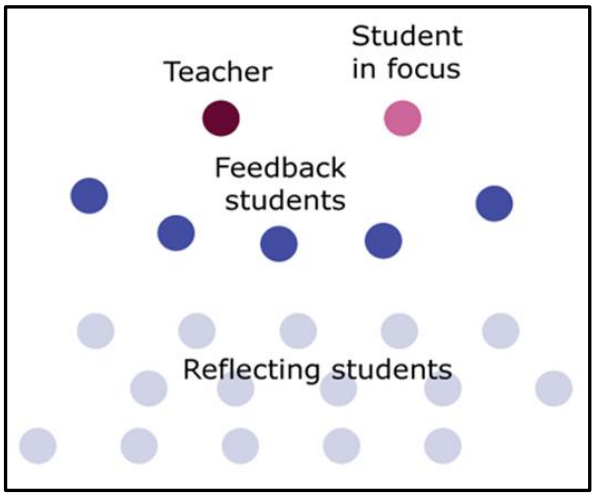


Figure 2. The three distinct phases of SAD and the classroom setting of students and the teacher according to their functional roles during the SAD.

From an assessment perspective, the SAD can provide evidence for what and how students are thinking. This information can be used for both formative and summative purposes. The latter because the SAD makes the level of students' understanding explicit for the teacher - and the

students are asked to assess themselves. From a formative perspective, the SAD holds prospects for activating students contributing to their own and peer students' learning and to voice their understanding so that the teacher can recognize and act on it to promote learning. Hence, one SAD session will lead to the formative and summative assessment of the focus student or focus group and possibly to formative assessment of many students.

Formative and summative assessment potentials associated with the SAD

The prospects for learning through formative assessment of a classroom dialogue may vary and can in the worst case be very minimal (Ruiz-Primo, 2011). The limited prospects for learning may be caused by multiple factors. The SAD intends to address four factors that each are expected to severely limit the prospects for learning. The first factor is related to insufficient planning, clarifying and sharing of learning intentions and criteria with students (Black & Wiliam, 2009; Hattie & Timperley, 2007; Ruiz-Primo, 2011). The second factor is associated with a limited engagement of students' in their own and peers' learning, including self-assessment and peer assessment (Black, Harrison, Lee, Marshall & Wiliam, 2004). The third factor is related to absence of, or unclear, feedback (Shute, 2008). The fourth factor is correlated to lack of alignment between goals, teaching and assessment approaches (Bennett, 2011; Krajcik, McNeill & Reiser, 2008).

In the following, we will elaborate on how SAD is designed to address the first three factors. Since learning goals describe the intended consequences of teaching and learning, they could form the basis for focusing and structuring the assessment. Based on this assumption, the SAD is guided by specific learning goals identified and described by the teacher ahead of the teaching and assessment session. A teacher may not make explicit – neither for him/herself nor for the students - the criteria for assessing whether, or at which level, a learning goal is being achieved. This will make it difficult for a student to recognize his/her own level or to provide feedback and engage actively in own and others' learning.

In the SAD, teachers are requested to subdivide the learning goals into a range of specific assessment criteria reflecting different aspects and levels of the competence being assessed. This is to avoid an unfocused assessment practice with a tendency to assess more general, trivial, or managerial aspects. Likewise, the purpose of this is to facilitate formative assessment. The rationale is that making different aspects and levels explicit will support more well founded decisions about next steps for individual students who may be at different stages in their learning and therefore requires different kinds of feedback. The subdivision of learning goals holds prospects for facilitating formative assessment by making the different aspects and levels in students learning process explicit. For summative purposes, the specific assessment criteria are expected not only to support the assessment of students' achievements, but also to provide transparency in grading.

In addition, teachers were asked to share and clarify to students the learning goals and range of criteria. The criteria are also used in the peer-feedback as well as in the self-assessment phase. It provides students with a tool to reflect on their current level of fulfilment of criteria aligned with the learning goals and on their next steps in learning. It is intended as a way of strengthening classroom learning cultures by having students engage actively with their own

and others' learning. Finally, providing the students with transparent assessment criteria would formalize the peer feedback, thus reduce the personal aspect among the students (e.g. friends, status).

In the next section, we will elaborate on how the fourth factor, related to lack of alignment, was addressed in the SAD design. According to Krajcik et al. (2008) consistency between goals and observable assessment criteria adapted to specific teaching sequences facilitates learning. In the same vein, Bennett (2011) argues for consistency between goals, teaching, and assessment approaches. Since teacher-led classroom dialogue is a very common instruction practice, a dialogue based assessment might feasibly facilitate consistency between instruction and assessment. In addition, a SAD session will typically be integrated in the ordinary teaching. Hence, the SAD will be situated within learning activities with which students have just engaged. Furthermore, students may be prompted to use artefacts (e.g. models, drawings, lab results) from the teaching during a SAD session. As aforementioned, the SAD is guided by learning goals described by the teacher ahead of the teaching and assessment session. This advanced clarification of learning goals might also facilitate consistency between instruction and assessment.

RESEARCH QUESTION

The purpose of the SAD is for students and teachers to gain insight into the students' current level of attainment and the next steps in students' learning in a useful and manageable way in the day-to-day teaching. The novelty of the SAD and teachers' central role in enacting the SAD guide our research question:

RQ: What are the main challenges and prospects perceived by teachers for using the SAD in the daily practice in science, technology and mathematics in lower and upper secondary schools?

RESEARCH METHODS

The SAD sessions were prepared by teams of teachers and researchers in Denmark and Finland, and the research was done in close collaboration with teachers as action research (Zeichner & Nofke, 2001). The SAD was implemented 20 times in Denmark and 6 times in Finland. Students were from lower and upper secondary school (level 7-12), and all teaching units had a focus on the cross-disciplinary competences modelling or argumentation.

The entire corpus of data consists of filled out teacher preparation templates, video recordings of the student-teacher dialogue, audio recordings of feedback sessions, filled out student self-reflection forms, an open-ended teacher questionnaire, and interview with teachers.

To answer the RQ, we used data from Denmark only. The data set consists of filled-in teacher preparation and reflections forms (n=11), two semi-structured interviews of teachers, one focus group interview with teachers (Kvale, 2007), and an open-ended questionnaire for teachers (n=4). The focus group interview and the semi-structured interviews were facilitated by two of the authors. Both the interviews and the responses to the open-ended questionnaire were analysed using thematic analysis (Braun & Clarke, 2006). We analysed the interview by

transcribing, reading, re-reading and in some cases re-listening to the interviews using the RQ1 as an analytical lens. Specifically, we focused on the challenges and benefits perceived by teachers with respect to each of the four principles behind the SAD. We used the same analytical approach for the questionnaire and the teacher preparation form.

RESULTS

Benefits related to planning, clarifying and sharing of learning intentions and criteria with students

The teachers saw several advantages in the formulation and utilisation of a clear statement of learning goals and explicit assessment criteria with respect to formative assessment, such as enhancing students' involvement in the assessment and providing transparency in the assessment. Teachers highlighted the benefits of sharing the assessment criteria with the students. A teacher wrote: "*The learning becomes explicit for the students*". Adding to this another teacher said: "*My students like it a lot and we will continue [*to use the method*]. Because they actually think that it worked - that the criteria were made clear and that they knew what to aim for*''. In addition, teachers found the criteria very useful for students to provide peer-feedback and self-reflection. Moreover, teachers' utterance also indicates how sharing of learning goals with students in the SAD becomes coherent with, and benefits, upcoming teaching by activating students in their own learning, exemplified by this quotation, "*I experienced that after a dialogue the students got better at setting up goals for themselves.*" In general, teachers acknowledge the SAD for facilitating students to take part in the assessment process. A teacher relates this to the following features of SAD: "*short and time-bound*.", and "*characterized by clear rules and roles*." According to the teacher those features enhance students' willingness to participate.

However, teachers not only appreciate SADs' features for facilitating students' engagement in their own learning but also features related to alignment between learning goal, teaching and assessment approaches.

Benefits related to alignment between learning goals, teaching and assessment approaches

In order to be useful for formative assessment and manageable in the day to day teaching, the assessment method must resemble and be easy to integrate into the ordinary teaching. As mentioned above, the usefulness with respect to formative assessment depends on the alignment between teaching and assessment approaches. In general, the teachers appreciated that SAD was dialogue-based and not written. This was mainly because the SAD in this way resembled the existing dialogue-based classroom practice, and at the same time provided teachers with a better prospect in understanding the basic ideas behind student's response. In this perspective a teacher highlighted the value of the SAD*:"[…..] and it is possible to get a nuanced picture of the students' understanding – something that a written text would not be able to capture to the same extent*."

With respect to manageability, teachers highlighted the SADs' adjustable features. As reflected in teachers' preparations and reflections forms and interviews, many teachers adapted the SAD

to local needs and contexts. E.g. the learning goals and their associated assessment criteria were adapted to specific teaching subjects and specific students' needs. In this way, teachers' use of the SAD assisted the process in aligning and adjusting the assessment with the ordinary teaching.

Teachers also adjusted the SAD to local class cultures. In classes with an ordinary teaching characterized by group work the single focus student was replaced by a group of two to four students. In classes with no need or tradition for grading, the SAD was only used for formative purposes. In this way, the adjustable feature of the SAD was used to strengthen the coherence between teaching and assessment. Furthermore, a teacher found that the restricted timeframe made the SAD a manageable assessment method to integrate in the relatively short timeframe of the regular teaching units.

Teachers use and acknowledgement of SAD to strengthen the alignment between the ordinary teaching and assessment was also extended to the subsequent teaching i.e. the assessment was used formatively to influence the following teaching. This is e.g. reflected in the teachers' descriptions of how the SAD facilitates students to take along their observations and reflections into the next teaching unit. The extended use of the SAD was also made by teachers. A teacher e.g. made a reference to how she followed up on a students' misconception, while another teacher made a reference to how the assessment criteria were expanded on in the subsequent teaching unit. This alignment between teaching and assessment was supported by appropriate teaching activities. For example, one teacher used SADs explicitly as a reference point to other activities: "*After the SAD-sessions, our students were to write a report based on the unit, and they could use the SAD there. They were motivated to use that as a shortcut to understanding how to present material*."

The SADs' prospects for strengthening the alignment between learning goals, teaching and assessment also relates to teachers' perceptions of SADs' prospects to address the assessment to different students.

Benefits related to aligning and adjusting to different students' learning needs, processes and achievements

Teachers experienced that the different aspect and levels reflected in the assessment criteria were useful in facilitating the learning process for a range of students with different learning abilities. This point is illustrated in this quote: "*During SAD, I mainly use the different assessment criteria for guiding my questions to address the differences between individual students' understanding and ability.*" This suggests that the teacher perceives the SAD as having potential for adapting the assessment criteria to different students. Note that this is even though the overall learning goal and its associated criteria are formulated in advance and targeted to assess students' achievements related to a specific teaching unit. In this way, teachers' use of questions based on the assessment criteria was used to strengthen the alignment between learning goals, teaching and assessment, and at the same time adjust the assessment to match different students.

Other adaptations used and appreciated by teachers were related to shortening or prolonging the total timeframe or single phases to fit the formative assessment to different students' needs

or give time for elaborating on new upcoming topics. This point is illustrated in this quote: "*And there was something they* [the students] *had a hard time to understand the last time, so if we had not gotten to that, the whole thing would have collapsed. So we took two more minutes*."

In the next section we will elaborate on how the teachers relate the promise of the SAD to the formative and summative purposes of assessment, respectively.

Benefits related to formative and summative purposes

Several of the benefits of SAD as perceived by teachers could be related to both summative and formative purposes. For instance, all the following aspects are related to usefulness when collecting evidence and judging students' achievements for assessment for summative as well as formative purposes, respectively: (1) clearly stated assessment criteria to enhance transparency and to guide questions and assessment; (2) alignment between learning goals, teaching and assessment; and (3) short and delimited in time and content, and (4) possible to get a nuanced picture of students' understanding and rationale through dialogue.

In general, however, the teachers acknowledged the SAD as a formative assessment method as illustrated in the following teacher quotes addressing the feature of the SAD: "*It captures the essence of formative assessment";* "*The main strength is the focus on formative assessment";* and "*It's very clear to the students that it is a part of a process."* This point is also reflected in the fact that in many utterances teachers do not only describe the SAD as an assessment method but as a *"teaching- and assessment tool"*.

Nevertheless, the teachers also saw several advantages in the formulation and utilisation of a clear statement of learning goals and explicit assessment criteria with respect to summative assessment, such as providing transparency in the assessment. A teacher used the learning goals for clarification, and for documentation for parents' meetings and to provide transparency in grading during the semi-annual student conversation.

As stated above, teachers described a large range of promise related the integrating the SAD in the ordinary teaching. However, teachers also encountered some challenges when enacting the SAD in their teaching and assessment practice.

Challenges related to planning and clarifying of learning expectations

Formative assessment is part of a learning process consisting of a series of learning steps, forming a progression. As part of the preparation for the SAD, teachers were asked to formulate learning goals with associated assessment criteria and questions to assess the different levels and aspects of students learning progress. To be explicit about the learning progression turned out to be one of the most challenging aspects for the teachers. In general teachers found it time consuming and difficult to prepare the learning goals, assessment criteria and questions: "*I think it has been time consuming to formulate different levels of assessment criteria", "I think there has been a lot of preparation; to sit down and really think through with assessment criteria and questions*", and *"It is not easy to make it clear for students what the criteria are*". All teachers emphasised that they value the formulation and use of the assessment criteria reflecting different levels and aspect of students learning progression. Still, many expressed

concerns with respect to being able to incorporate this very time consuming preparation in the day to day teaching and assessment practice.

Even with the operationalization of the learning progression into assessment criteria, teachers found it challenging to judge the level of the focus student: "*It is not easy to work with learning progressions and planning for giving summative assessment at the end. How can I in five minutes of dialogue and five minutes of feedback be sure that someone asks questions, which will allow me to place the student on one of the progression steps?* "

Another challenge raised by teachers relates to striking an appropriate balance between knowing what the focus student is capable of and which questions to include in the dialogue and, at the same time, clarifying a realistic level of learning expectations to the rest of the class. One teacher described, "*It* [the task and questions] *must resemble the appropriate complexity required in a teaching situation and for the final exam. It should not be too easy* […]. *You have to find the right student to deal with that* [the complexity]. *But it's not an easy task to strike the balance."* Another concern voiced by the teachers related to this issue was to avoid display of weak students' level of achievements in front of their peers.

This confirms that an important part of planning is for the teacher to tailor the questions to the focus-student while still making realistic assessment criteria clear to other students. Adding to this, the teacher must ensure his/her questioning of the focus-student provides other students with sufficient information enabling them to provide sound peer-feedback.

Challenges related to students' involvement in their own and others learning

In general, teachers perceived the peer-feedback session as the most challenging part of the SAD due to students' inadequate "assessment literacy", such as low assessment value with respect to both feedback quantity and quality. E.g. a teacher wrote, "*In the peer-feedback session I missed content depth and more comments*". Teachers mainly addressed challenges related to assessment literacy with respect to students' limited content knowledge and praised their peers instead of providing guidance for the next step in the learning. To address this challenge, the teachers often perceived a need to add to or facilitate the peer-feedback session. A teacher was planning to repeat the SAD but nuance it in the following way: "*Allocating different roles to the students in the feedback group, so that each student gets his/her own assignment*". Another teacher was planning to provide the students with a rubric with the learning goals divided into three different levels of learning progressions.

Finally, teachers encountered challenges related to activating all students. Teachers described that students' (but not all and with variation in their effort) took an active involvement in the process. However, teachers also reported SAD sessions where it was hard to activate all students throughout the session: "*The listening students may have a hard time keeping up*" and "*It is a challenge to keep the drive-over-time in the SAD so that the feedback group is serious about their own learning* (self-assessment)."

Challenges related to time frame and alignment between teaching and assessment approaches

The SADs were integrated as a part of the ordinary teaching to make alignment between teaching and assessment. The strict 5-minute time frame in the dialogue posed a challenge to most teachers to assess students' achievements with respect to the previous teaching. The teachers describe how the time frame limits the amount and complexity of the content to be assessed. As one teacher made explicit: "*It can be difficult to make as limited an assessment that it is possible to keep the time frame*". A different teacher states: "*After five minutes we were just started. I was not able to address modelling appropriately".* Based on experiences from an implemented SAD, the same teacher expressed how he adjusted the next SAD to the restricted timeframe: "*The more complex questions were toned down.*" Another teacher was also planning to repeat the SAD but adjusted the timeframe instead of the complexity: "*I probably would not obey the five minutes, but use the time that is needed on the dialogue."*

Some teachers chose a group of students to be in focus rather than just one focus student. This was done in order to make alignment to and resemble their current classroom practice (i.e. group work), moderate the feeling of high-stakes assessment and avoiding exposing a single student. However, this made the five-minute timeframe even more challenging.

Combining summative and formative assessment

Most teachers used the SAD for only formative purposes, meaning that the possibility of combining the two uses of assessment was not so well examined.

Regarding students' self-summative assessment, teachers in general, but with exceptions, believed that the students assess themselves on too high a level of achievement. A teacher wrote "*No demands for giving feedback to students self-assessment as I doubt the function, validity and seriousness.*"

During the peer-feedback, which had mainly a formative purpose, there was a tendency only to comment on positive aspects: "*When female friends were feedback students, they only provide each other positive feedback".* A teacher believed that the "over grading" in self-assessment and the insufficient feedback were part of a "*performance culture*". This point is illustrated in the following quotes: *"Students' didn't believe me when I told them that it (*the SAD*) was a kind of a play and that it would not influence the grading at all."* and "*They think I will look at the* (self) *grade and base my grading on it."* Another challenge highlighted by teachers is related to the SAD's physical set up: "*There is a tendency that students may experience the SAD as an interrogation. When that is the case, the students will not see the process as being useful with a view to the future."*

Due to the strictness of the formative processes based on a clear and explicit learning progression in the competences assessed, the SAD has a potential for summative use without distorting the formative aspects. This is because that any summative assessment happens at the very end of the ritual and is embedded in a formative process. We thus avoid the before mentioned domination of the summative purpose; you tend to see when the two purposes are mixed (Butler, 1987).

But it was clear that the performance culture in Danish schools made the summative use a delicate thing and the tendency to perceive the SAD as a kind of exam might hinder the formative prospects in the SAD.

Future perspectives on research on structured assessment dialogues

For the purposes of this paper, we have focused exclusively on teachers' perceptions of the SAD. Since teachers' and students' perceptions of feedback can be different (Ellegaard et al, 2017), a future study of student perceptions of the SAD might provide insights into why, for example, the feedback part of the SAD is difficult to orchestrate. Our current corpus of data will not allow us to investigate student perceptions directly.

Another future perspective is to investigate the actual dialogues and their role in relation to teaching and learning in the science classroom. In other publications, we have developed a methodology for mapping and analyzing the SAD (Dolin et al, 2018b). The mapping we have developed integrates network analysis with a dialogical coding scheme, criteria for the dialogue, and gestures. Each dialogue is then converted into a map, which shows who is active, which criteria are addressed, and which dialogical strategies are used in the dialogue. We have used this methodology to extract a typology of dialogues. In future studies, such a typology could be used to characterize the role of different kinds of the SAD in the science classroom.

CONCLUSIONS

In most of the enactments of the SAD, teachers only used the formative potentials of the SAD. The formative prospects are mainly related to the SADs' features of: (1) clarifying and sharing assessment criteria with students; (2) enabling high student engagement in their own and peers' learning; and (3) being adjustable to students with different needs.

In addition, teachers acknowledged the SADs potential for both formative and summative purposes related to SADs' features of: (1) facilitating coherence between teaching and assessment approaches; (2) being adjustable in time and content; and (3) being relatively easy to enact and integrate into the existing teaching practice.

The SAD has potential for combining formative and summative purposes of assessment. However, it needs careful and repeated enacting if the summative aspects should not hinder the formative prospects for instance through change of performance classroom culture, and through enhancing and supporting students' competences in providing feedback and self-grading. Finally, teachers need time and experience in identifying and describing appropriate assessment criteria that reflect different levels and aspects of students' learning process.

REFERENCES

Bakhtin, M. M. (1981). *The Dialogic Imagination*. Austin: University of Texas Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working Inside the Black Box: Assessment for Learning in the Classroom. (Cover story). *Phi Delta Kappan*, *86*(1), 9–21.

Black, P., & Wiliam, D. (1998). Developing a theory of formative assessment. In: Gardener, J. (Ed.): *Assessment and Learning* (pp. 81–100). London: Sage.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and involvement. *British Journal of Educational Psychology*, *58*, 1–14.

Dolin, J., Black, P., Harlen, W., & Tiberghien, A. (2018a). Exploring relations between formative and summative assessment. In: Dolin, J & Evans, B. (Eds.): *Transforming assessment: Through an interplay between practice, research and policy* (pp. 109–140). Cham, Switzerland: Springer.

Dolin, J., Bruun, J., Nielsen, S. S., Jensen, S. B., & Nieminen, P. (2018b). The structured assessment dialogue. In: Dolin, J & Evans, B. (Eds.): *Transforming assessment: Through an interplay between practice, research and policy* (pp. 109–140). Cham, Switzerland: Springer.

Dysthe, O. (1996). The multivoiced classroom interactions of writing and classroom discourse. *Written Communication*, *13*(3), 385–425.

Ellegaard, M., Damsgaard, L., Bruun, J., & Johannsen, B. F. (2017). Patterns in the form of formative feedback and student response. *Assessment & Evaluation in Higher Education*, 1-18, doi.org/10.1080/02602938.2017.1403564

Harlen, W. (2012). On the relationship between assessment for formative and summative purposes. In: Gardner, J., (Ed.), *Assessment and Learning* (pp. 87–102). London: Sage

Harlen, W. (2013). *Assessment & inquiry-based science education: Issues in policy and practice*. Trieste: Global Network of Science Academies (IAP) Science Education Programme (SEP). http://www.interacademies.net/File.aspx?id=21245 (retrieved 12.01.2017).

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education*, *92*, 1–32.

Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation*, *37*(1), 15–24.

Shinn, M. R. (2013). *Measuring general outcomes: A critical component in scientific and practical progress monitoring practices*. Pearson: Aimsweb. Http://www.aimsweb.com/Wp-content/uploads/Mark-Shinn-gom_Master-Monitoring-White-paper.pdf (retrieved 12.01.2017).

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.

Wiliam, D. (2011). *Embedded Formative Assessment, Study Guide*. Bloomington, IN: Solution Tree Press.

Wiliam, D., & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for F-12 classrooms*. Moorabbin VIC: Hawker Brownlow Education.

Zeichner, K. M., & Nofke, S. E. (2001). Practitioner research. In V. Richardson (Ed.), *Handbook of Research on Teaching* (4th edition). Washington DC: Aera.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

AN INSTRUMENT FOR MEASURING PUPILS' FAMILIARITY WITH SCIENCE EDUCATION SETTINGS

Rebecca Cors[1], Andreas Müller[2] and Nicolas Robin[1]
[1,3]University of Teacher Education, Institute for Teaching Natural Science, St. Gallen, Switzerland
[2] Universite de Geneve, Fac. of Science/ Physics Sect., and Institute of Teacher Education, Geneva, Switzerland

How novel, or unfamiliar, a visit to an out-of-school learning place (OSLeP), such as a science center, feels has been linked to changes in young people's interest in science and technology (S&T). However, few studies have attempted to measure how participants perceive novelty during OSLeP visits. This article describes the development, testing, and validation of a survey instrument to measure at-visit perceptions of novelty at OSLePs. Drawing from existing studies about how people perceive novelty in learning settings, researchers developed a questionnaire to assess at-visit perceptions of novelty. A total of 215 pupils completed the survey during a mobile laboratory visit (www.mobillab.ch). Through factor analysis and reliability testing, the authors identified four meaningful clusters of survey items. These survey scales, which we called novelty experience factors (NEFs), define four dimensions of pupils' at-visit perceived novelty: curiosity (state), exploratory behavior, oriented feeling, and cognitive load. Results also offer new insights into how pupils perceived novelty at OSLePs. These NEFs were useful in a larger study that investigated relations between pupil factors, at-visit novelty, and educational outcomes for an informal science learning program. Measuring at-visit novelty is a key to better understanding the effectiveness of OSLeP experiences for promoting educational outcomes.

Keywords: science interest, novelty, out-of-school learning

INTRODUCTION

Over the last forty years, out-of-school learning places (OSLePs), such as science centers and mobile laboratories, have been developed to promote interest in science and technology (S&T) topics and careers. These programs are critical for educating our Digital Age workforce and for promoting a scientifically literate citizenry (Sjøberg & Schreiner, 2010).

Novelty is a key factor for investigation of informal science education programs

How novel, or unfamiliar, an OSLeP visit feels to people has been shown to affect the degree to which their dispositional, or lasting, interest in the topic of the OSLeP develops. For example, faced with operating unfamiliar, high-technology equipment, some people feel intrigued, while others feel intimidated by the endeavor.

Measures of at-visit novelty are lacking

Even though studies of novelty at OSLePs suggest that a better understanding of at-visit novelty is important, few have measured it. The authors reviewed eight studies that examine how feelings of unfamiliarity, or perceived novelty, related to learner knowledge gain and S&T interest development at OSLePs (Anderson & Lucas, 1997; Cors et al., 2015; Cotton & Cotton, 2009; Falk & Balling, 1982; Falk et al., 1978; Jarvis & Pell, 2005; Kubota & Olstad, 1991;

Orion & Hofstein, 1991). Only three of these studies examined at-visit, perceived novelty at OSLePs, measured as exploratory behavior (Falk & Balling, 1982; Falk et al., 1978; Kubota & Olstad, 1991). Two of the studies showed that learners who were supposed to have more familiarity with the setting, either because they lived near a setting similar to the OSLeP (Falk et al., 1978) or because they saw a pre-visit orienting video (Kubota & Olstad, 1991), exhibited more exploratory behavior. Falk and Balling (1982) also found significant links between learning setting and exploratory behavior, which varied based on learner age. They found that third graders carrying out an assignment at a nature center, a more novel setting, worked less 'on-task' and displayed more social discomfort than their peers who did the assignment in a familiar wooded area next to their school. In contrast, fifth graders doing the same assignment in a wooded area near their school, a less novel setting, worked less 'on-task' and showed more signs of boredom than their peers who were working at the nature center.

That several studies about OSLePs showed links between exploratory behavior and learning setting highlights the importance of understanding at-visit novelty when investigating the effectiveness of OSLePs. For this reason, the authors wanted to measure at-visit novelty as part of their study of a science education program in Eastern Switzerland called mobiLLab (mobiLLab.ch). The mobiLLab program was developed by faculty at the University of Teacher Education in St. Gallen, Switzerland. Each semester, staff deliver 12 experimental stations to schools so that pupils can have a half-day experience with equipment used at S&T industries in the area. Indeed, a pilot study had already elicited teacher comments indicating that the pupils' comfort and familiarity with mobiLLab experimental equipment improved their ability to profit from a mobiLLab visit (Cors et al., 2015).

Clues for measuring at-visit novelty

The studies included in the literature review describe several indicators of at-visit novelty that have not been explored for studies of novelty at OSLePs, but that have been explored through related research. For example, several researchers describe pre-visit 'orienting' activities that should have reduced novelty, yet none measure the extent to which learners feel oriented at the OSLeP visit (Anderson & Lucas, 1997; Jarvis & Pell, 2005; Kubota & Olstad, 1991; Orion & Hofstein, 1991). However, in a related study, a parameter similar to oriented feeling, called 'preparation and orientation,' was linked to pre-visit classroom preparation activities before an OSLeP visit (Orion et al., 1997). Another unexplored indicator of perceived novelty named by previous studies of OSLePs is at-visit curiosity (Anderson & Lucas, 1997). Such situational curiosity has, in turn, been related to the diversive and specific components of epistemic curiosity (Litman & Spielberger, 2003). A third unexplored indicator of at-visit novelty is described by some studies as how overwhelmed by unfamiliarity learners are at by experiences at OSLePs (Falk et al., 1978; Kubota & Olstad, 1991). This feeling of being overwhelmed by new information or objects could be measured using cognitive load survey scales that have been developed for flight simulator training settings (Hart & Staveland, 1988).

This existing research provided a foundation of survey items that could be adapted to measure at-visit novelty as perceived by visitors at OSLePs. This paper describes how the authors developed, tested and validated a survey instrument to measure pupils' perceived novelty during a mobiLLab visit.

METHODS

Drawing from this research, a 20-item questionnaire about pupils' at-visit novelty was developed, pilot tested, and then distributed to more than 200 pupils who visited mobiLLab, a science education laboratory in St. Gallen, Switzerland, in 2015. A factor analysis and reliability testing were used to identify measures for four dimensions of at-visit perceived novelty. We refer to these dimensions of perceived at visit novelty as novelty experience factors (NEFs).

Instrument Development and Testing

The first step in developing an instrument to measure pupils' at-visit novelty experience was to look at survey items developed for similar purposes. Twenty survey items were adapted from other studies about how oriented, curious, free to explore, and overloaded people feel in a learning setting.

Piloting of the survey took place during a mobiLLab school visit in December 2014, where 40 pupils completed the survey during a break after they worked through their first two experimental stations and before they worked through their two remaining stations. For most items, pupils were asked to mark their level of agreement with each item on a 4-point Likert scale ranging from 'not at all true' to 'completely true.' An example item for curiosity is 'I would like to learn more about the mobiLLab science themes and topics.' The workload items had a 4-point semantic differential scale. For example, the item 'What did you think of the time allotted to carry out the experiments?' had endpoints of 'too long' and 'too short' to describe visit length.

After reviewing the variation in scalar responses and written responses, we revised several survey items. Also, because teacher feedback indicated that having the students complete the surveys during the break in the mobiLLab visit worked well, that schedule was maintained.

Data collection

A total of 215 pupils in 21 different class groups at 7 schools completed the at-visit survey during mobiLLab visits during the spring of 2015. Pupils were aged 13–15 and attended a secondary school in Eastern Switzerland that would prepare them for a trade or vocational program.

Validation: factor analysis approach

Because the survey items were adapted from several different previous studies and used in a new combination, we needed to explore how the items would describe different dimensions of pupils' novelty experience at the mobiLLab visit in particular. It was important to distinguish those items that elicited responses indicating perceptions of, for example, curiosity state, from, for example, those items that characterized exploratory behavior. By revealing how responses to certain survey items have common variance, exploratory factor analysis helps researchers to identify clusters of items, or scales, which reliably represent a characteristic about a population. For these reasons, we chose principal axis factoring (Field, 2013) to explore the items about how pupils perceived novelty at the mobiLLab visit.

Through exploratory factor analysis, we looked at how responses to certain groups of survey items explained the variance in pupil responses to the entire survey in a similar way, indicated by a factor loading of > .03. A factor loading indicates the relative contribution of an item to a factor. It can be thought of as the Pearson correlation coefficient between a factor and a survey item. For example, the item, 'I would like to learn more about the mobiLLab science themes and topics,' had the relatively high loading of .67 for the factor called curiosity state. This high loading shows that the survey item strongly contributes to the variation among the items of the curiosity state factor. By reviewing these survey items and examining their relatability as a group, we sought to identify groups of items that reliably represented pupils' at-visit novelty experience.

RESULTS

Results show that exploratory factor analysis and reliability testing were useful ways to examine a group of survey items about perceived novelty that had not yet been used together. It enabled us to determine that there were several explanatory factors under which the items can be grouped as viable survey scales (measurement instruments), namely curiosity state, exploratory behavior and cognitive load.

First, an initial principal analysis factoring, with orthogonal (Varimax) rotation, was run with the 20 survey items, which had been adapted from other studies. The analysis produced an eigenvalue for each factor. Eigenvalues represent the amount of variation among all survey items in the questionnaire that can be explained by that factor. Four factors had eigenvalues greater than 1.1 and these same four factors were identified by oblique (Oblimin and Promax) rotations. A fifth factor also had an eigenvalue greater than 1, but was eliminated because it included only one item with factor loadings greater than 0.3. The four factors in combination explained 47% of the variance in pupil responses. The scree plot was somewhat ambiguous and the inflection supported selecting either three or four factors. Given the moderate sample size and confirmation by Oblimin and Promax rotations, we retained three factors. The items that clustered on the same factor suggested that factor 1 represented pupils' curiosity state, factor 2 represented pupils' cognitive load, and factor 3 represented pupils' exploratory behavior with mobiLLab equipment. A fourth factor, representing the extent to which pupils felt oriented, showed only two items with factor loadings greater than 0.3, which together gave a Cronbach's alpha of α=.53, too low to warrant their use as a reliable scale. However, one of the items loaded at lambda=.69, so this item was used alone to represent at-visit oriented feeling.

Next, we conducted a forced three-factor analysis to identify which survey items contributed to scales for the three remaining NEFs: curiosity state, cognitive load, and exploratory behavior. Cronbach's alphas for these three groups helped determine that two items were not contributing to reliability of the three strong factors. The item cls2 had a loading of < .300, so it was eliminated. Cronbach's alpha for the exploratory behavior scale turned out to be greater than .700 when exex1 was eliminated and texs4 was used in reverse form.

A final principal axis was run as a Varimax rotation with the remaining18 items (excluding cls2 and exex1). The results are shown in **Error! Reference source not found.**, which lists

factor loading and communalities for each item and lists Cronbach's alphas and eigenvalues for each factor. Through this final test (N=205), three factors were identified through loadings and only these factors had eigenvalues greater than 1. By reviewing the item clustering that this last factor analysis revealed, and using common sense and reliability testing, we identified survey scales for three dimensions of at-visit novelty at OSLePs. The dimensions are curiosity state, which explains 29% of the variance in pupils' responses to all of the survey items, exploratory behavior, which explains 12% of the variance, and cognitive load, which explains 8% of the variance.

Error! Reference source not found. provides a summary of the variables developed as a result of this study. For each NEF, the table lists the number of survey items, the percent variance each factor explains, and the reliability for each scale.

Table 1: Number of survey items, percent variance, and reliability for four Novelty Experience Factors (NEFs) (N=205).

Novelty Experience Factor (NEFs)	Number of survey items	Percent Variance in Pupil Responses	Reliability: Cronbach's α
Curiosity State	6	29%	.86
Exploratory Behavior	7	12%	.70
Cognitive Load	6	8%	.70
Oriented Feeling	1	NA	NA

DISCUSSION

This study produced and tested four measures of perceived novelty at OSLePs, all of which are grounded in novelty theory. The results also provide the first psychometric validation for three measures of perceived novelty at OLSePs. That is, investigators can use these survey scales to measure pupils' perceived novelty in terms of reported curiosity state, exploratory behavior and cognitive load. While a variable for a fourth measure of perceived novelty, oriented feeling, was identified, it was a single survey item, rather than a scale.

By looking more closely at the factor loadings from factor analysis results, one gains further insight into ways that pupils perceived novelty during their visit. First, pupils associate the feeling of being oriented with the feeling of being able to explore the equipment. Evidence of this is that items sett1 and sett2 and sett3, which were developed to describe feeling oriented, loaded for the exploratory behavior factor. Also, pupils associated cognitive load with conducting experiments, as seen by the fact that items exex2 and exex3, which measure pupils' ease with experimenting, loaded on the cognitive load factor. Finally, the fact that item texs5 belongs not to the exploratory behavior scale, as expected, but to the curiosity scale suggests that pupils associate fun more strongly with curiosity than with exploratory behavior.

Table 2: Results of the factor analysis.

Item	Factor			ĥ
	Curiosity State	Exploratory Behavior	Cognitive Load	
curs1: Die Erfahrung mit mobiLLab weckt meine Neugier auf die dort behandelten Themen.	**0.78**	<.30	<.30	.659
curs2: Es interessiert mich, wie die Geräte an den verschiedenen Posten funktionieren.	**0.70**	<.30	<.30	.527
curs5: Ich möchte die in den mobiLLab behandelten Themen besser verstehen.	**0.70**	<.30	<.30	.522
curs4: Die in den mobiLLab-Versuchen behandelten Themen haben mich persönlich angesprochen.	**0.69**	<.30	<.30	.558
curs3: Ich möchte mehr über die mobiLLab-Themen erfahren.	**0.67**	<.30	<.30	.474
texs5: Es hat mir Spass gemacht, die mobiLLab-Geräte auszuprobieren.	**0.59**	<.30	<.30	.457
texs1: Ich habe keine Probleme, die mobiLLab-Geräte selbst zu bedienen.	<.30	**0.51**	<.30	.355
setts3: Für den mobiLLab-Besuch bin ich gut vorbereitet.	<.30	**0.48**	<.30	.272
texs4: Ich konnte rasch mit der Bedienung der mobiLLab-Geräte beginnen.	<.30	**0.47**	**0.33**	.342
texs2: Aufgrund der Vorbereitung habe ich keine Angst, bei der Bedienung der mobiLLab-Geräte Fehler zu machen.	<.30	**0.46**	<.30	.232
setts1: Der zeitliche Ablauf des mobiLLab-Tages ist mir bekannt.	<.30	**0.46**	<.30	.286
setts2: Der mobiLLab-Besuch ist gut organisiert.	<.30	**0.43**	<.30	.265
texs3: Ich bin in der Lage mit den mobiLLab-Geräten zu „spielen“ um zu sehen, was sie alles können.	<.30	**0.39**	<.30	.262
cls3: Wie sehr musstest du dich anstrengen, um die Experimente durchzuführen?	<.30	<.30	**-0.53**	.290
cls1: Wie hoch war die geistige Belastung bei den Versuchen insgesamt (zuviel Unbekanntes, zuviel auf einmal)?	<.30	<.30	**-0.52**	.275
exex3: Ich konnte mich gut auf die Experimente konzentrieren, ohne mit den Geräten “kämpfen” zu müssen.	<.30	0.34	**0.52**	.431
cls4: Wie verunsichert, entmutigt, oder verärgert warst du während der Experimente?	<.30	<.30	**-0.47**	.326
exex2: Die Experimente waren schwierig.	<.30	<.30	**-0.45**	.288
cls2: Wie empfindest du die Zeit, die für Experimente zur Verfügung stand?	NA	NA	NA	NA
exex1: Wir haben genügend Informationen, um die Experimente durchführen zu können.	NA	NA	NA	NA
Cronbach’s α	0.86	0.70	0.70	
Eigenvalue Total	5.21	2.11	1.36	
% of Variance	28.93	11.73	7.55	
Cumulative Variance	28.93	40.65	48.21	

Measuring at-visit novelty can help researchers and educators to better understand the degree to which learners perceive their OSLeP experience as new or unfamiliar. For example, a measure of oriented feeling can indicate whether, and by how much, novelty was reduced by classroom preparation. Such data can help us untangle the effects of classroom preparation from the many other variables that affect learner experiences at OSLePs. Measuring at-visit novelty is key to developing strategies for leveraging both 'negative' novelty (cognitive load), and 'positive' novelty (curiosity state, exploratory behavior, oriented feeling), in order to promote the effectiveness of OSLePs.

REFERENCES

Anderson, D., & Lucas, K. (1997). The Effectiveness of Orienting Students to the Physical Features of a Science Museum Prior to Visitation. *Research in Science Education, 27*(4), 485-495.

Cors, R., Müller, A., & Robin, N. (2015). Advancing Informal MINT Learning: Preparation and Novelty at a Mobile Laboratory. *New Perspectives in Science Education*, 53-58.

Cors, R., Müller, A., & Robin, N. (2016). A Informal Science Learning: an investigation of how novelty and motivation affect interest development at a mobile laboratory. University of Geneva (Thesis)

Cotton, D. R. E., & Cotton, P. (2009). Field biology experiences of undergraduate students: the impact of novelty space. *Journal of Biology Education, 43*(4), 169-174.

Falk, J. H., & Balling, J. D. (1982). The Field Trip Milieu: Learning and Behavior as a Function of Contextual Events. *Journal of Educational Research, 76*(1), 22-28.

Falk, J. H., Martin, W. M., & Balling, J. D. (1978). The Novel Field-Trip Phenomenon: Adjustment to Novel Settings Interferes with Task Learning. *Journal of Research in Science Teaching, 15*(2), 127-134.

Field, A. (2013). Discovering Statistics Using IBM SPSS Statistics. London: SAGE publications.

Hart, S., & Staveland, L. (1988). Development of the NASA-TLX (Task Load Index). In P. A. H. N. Meshkati (Ed.), Human Mental Workload (pp. 239-250). Amsterdam: North Holland Press.

Jarvis, T., & Pell, A. (2005). Factors Influencing Elementary School Children's Attitudes toward Science before, during, and after a Visit to the UK National Space Centre. *Journal of Research in Science Teaching, 42*(1), 53-83.

Kubota, C. A., & Olstad, R. G. (1991). Effects of Novelty-Reducing Preparation on Exploratory Behavior and Cognitive Learning in a Science Museum Setting. *Journal of Research in Science Teaching, 28*(3), 225–234.

Litman, J. A., & Spielberger, C. D. (2003). Measuring Epistemic Curiosity and its Diversive and Specific Components. *Journal of Personality Assessment, 80*(1), 75-86.

Orion, N., & Hofstein, A. (1991). Factors which influence learning ability during a scientific field trip in a natural environment. Paper presented at the Proceedings of the annual convention of the National Association for research in Science Teaching, Fontana, IL.

Orion, N., Hofstein, A., Tamir, P., & Giddings, G. (1997). Development and Validation of an Instrument for Assessment the Learning Environment of Outdoor Science Activities. *Science Education, 81*, 161-171.

Sjøberg, S., & Schreiner, C. (2010). The ROSE project: An overview and key findings (pp. 31): University of Oslo.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

FRAMING DOCTORAL SUPERVISION AS FORMATIVE ASSESSMENT

Sofie Kobayashi
Department of Science Education, University of Copenhagen, Copenhagen, Denmark

This paper addresses the issue of developing autonomy in PhD education by drawing from and integrating two separate research domains within PhD education research: One domain is research that explores how supervisors can support the development of autonomy. The other domain is research into assessment criteria for the PhD. The two domains are integrated in a model of learning through formative assessment, which is translated into the realm of PhD education. The aim is to help supervisors enhancing the learning and competence development of PhD students and supporting their autonomy. The model builds on the use of explicit assessment criteria and involves PhD students assessing their own work. The model should always be adapted to the concrete domain and practices of each PhD study. Further research is suggested to uncover and develop explicit assessment criteria that are discipline specific in science and sufficiently detailed to be operational for supervisors and PhD students.

Keywords: PhD supervision, autonomy, formative assessment

SETTING THE SCENE

"they enter a domain where it is very difficult to measure success criteria. The things we think are important, independence and those things, it's kind of difficult to measure" (Asger, PhD supervisor)

"I think I did a very good job, but finally, a lot of comments stem from that part. That means that is not a good job. That's kind of problem that confuses me a lot. It's very... maybe I don't have that good competence to evaluate [my] own work." (Wang, PhD student)

These two quotations from interviews with PhD students and supervisors set the scene for the model I bring into PhD education in this article. The supervisor points to the difficulty in explicating assessment criteria for PhD education, as the aim of PhD education is to produce independent or autonomous researchers. What is assessed in the examination process is first and foremost the thesis, which should document the PhD student's ability to produce new knowledge formulated as original or a (significant) contribution to science (Tinkler & Jackson, 2004). On the other hand the PhD student points to one way of recognising autonomy; to be able to assess one's own work. While we know that assessment and self-assessment requires criteria, both supervisors and PhD students would benefit from an overall framework to understand and overview the learning process in PhD supervision with emphasis on the use of criteria. Research into assessment indicates that self-regulation and hence autonomy can be supported through involvement of students in the assessment process (Boud & Soler, 2015; Dolin et al., 2017), but the link to doctoral supervision has not been made so far. The central position students get in assessing their own work help them understand their own learning process, and work towards the goals and standards of the discipline. At a higher level of learning they build competences in assessing their own work beyond the timescale of the course

or studies. To assist supervisors and PhD students in this, I bring in and adjust a model for formative assessment in teaching at undergraduate level developed by Dolin et al. (2017).

Intensions

The aim is to suggest a model of the learning process in the context of PhD education and supervision that can help supervisors in enhancing the learning and competence development of PhD students and supporting their autonomy. The model involves PhD students in assessing their own work through clear and explicit criteria. The model is a further development of the model of formative assessment developed by Dolin et al. (2017).

To adjust the model to the realm of PhD education I draw on two separate research domains within PhD education research. One domain is research that explores how supervisors can support the development of autonomy. The other domain is research into assessment criteria for the PhD. I integrate these two research domains into a model for formative assessment. The model has its limitations in that there surely are learning processes that cannot be captured in this model and the model should always be adapted to the concrete domain and practices of each PhD study.

RESEARCH INTO AUTONOMY AND ASSESSMENT CRITERIA

Research into supporting autonomy in PhD education

New doctoral supervisors often point to building PhD student autonomy as the most difficult and pressing issue in their development as supervisors. The challenges that supervisors state in my workshops are for instance: "*Finding the right balance between facilitating and controlling the process*", "*To strike the right balance between reactiveness and pro-activeness - when to push/pull or nurse - and when to ´wait´ and give time for the student to show up own initiative and own work*", or "*Finding the right level of supervision (guidance versus independence)*". Delamont, Parry and Atkinson (1998) vividly describe how supervisors experience the difficulties in creating this delicate balance, while Gardner (2008) describes the dilemma from the students' point of view.

In his much cited work Gurr (2001) refers to the two styles of supervision as 'hands-on' and 'hands-off'. The assumption is that PhD students need more direction and hands-on guidance when they are dependent, and more 'hands-off' supervision as they become competent autonomous. I will return to his term 'competence autonomous' later. Gurr does not discuss exactly what supervisors can do to support autonomy, but his toolkit is suggested as a way to open discussions about this between supervisor and PhD student, and as such it is a tool to initiate meta-communication about supervision.

While Gurr depicts a rough development from dependency to autonomy over time, Kam (1997) differentiates between three dimensions of supervisor dependency: 'work organisation and problem solving', 'research preparation' and 'communication', and describes the specific needs that students in each category voice. Such differentiation is useful in helping supervisors to meet the needs of their students.

However, there is an aspect of supervision that it is important to consider when meeting students' needs - the importance of shared intensions and meta-communication in empowering

the learner. If a supervisor primarily attends to the needs of a student without meta-communicating about the intensions of the help they provide, then there is a danger that the student will remain in need of help (Molly & Kobayashi, 2014; Strong et al., 2008). Meta-communication about intensions is a way of putting students in charge of their own learning process (Baltzersen, 2013) and especially when students are supported in making their own decisions the transparency will empower students and support autonomy. The model described by Strong et al. (2008) builds on Karl Tomm's postures in collaborative counselling as shown in Figure 1. Empowerment is facilitated by shared intensions (transparency) and an open decision space.

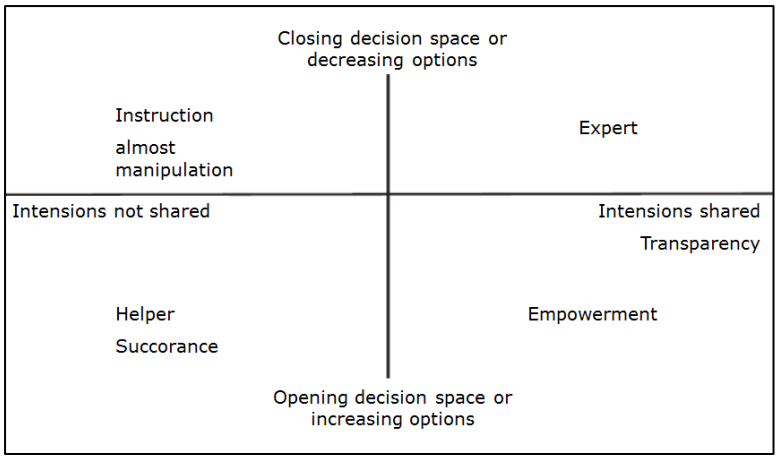


Figure 1. Karl Tomm's Collaborative Approaches to Counselling, adapted from Strong et al. (2008).

This is supported by the research conducted by Overall, Deane and Peterson (2011). They use research self-efficacy as an indicator of autonomy, defined as 'how much the student believes s/he can successfully complete key tasks, such as data collection, data analyses and writing articles' (p. 792). They then investigate how effective different types of supervisor support are in building student research self-efficacy beliefs, and they find that autonomy support combined with task-related academic support is most efficient. They do not investigate the effects of meta-communication or sharing intensions, but parts of that may be deducted from their items (p. 793-94). The autonomy support seems to consist of mainly meta-communication, e.g. 'My supervisor encourages me to ask questions'. The task-related academic support has few items that directly suggest meta-communication, but the first item does: 'My supervisor provides clear expectations and goals I need to achieve'. The rest of the items seem to be more expert advice, but combined with the open decision space it suggests that the style of supervision (or approaches to counselling) that Strong et al. (2008) term Empowerment. It could be interesting to investigate the importance of 'providing clear expectations and goals' since goals and assessment criteria play an important role in assessment.

Lovitts (2005, 2008) point to a number of components that she identifies as critical for doctoral students to make the transition from being dependent on close supervision to the stage where they are expected to be autonomous (independent) researchers. The independent researcher is seen as capable of making an original contribution to knowledge, which she argues requires creative performance. She identifies self-direction, perseverance, tolerance of ambiguity, a

willingness to take risks and intrinsic motivation as factors that are critical to creative performance, and hence completion. However, she does not suggest what supervisors can do to enhance these components, except for changing their behaviours to 'better enhance and support the development of the subcomponents identified as most critical to creative performance' (p. 151).

While Lovitts (2005) link autonomy and independence to creativity, other scholars emphasise the link to critical thinking and the ability to evaluate one's own work. This is the perception that the PhD student voices in the opening of this article. Holbrook et al. (2004) find that 'Examiners expected a balanced and critical appraisal of both the literature and the candidate's own findings…' (p. 112), but references to critical thinking as an intended learning outcome of the PhD process are rare and it is rather an implicit assumption in assessment, as seen in Tinkler and Jackson (2004) and also observed by Brodin (2016). It seems to be left to the examiners to be critical. One can only speculate on the lacking reference in the explicit learning outcomes on critical thinking and the ability to assess own work. What is assessed (in the first place) is the thesis, and this is a summative assessment judging whether the thesis reflects originality or a contribution to science, whether amendments are needed or whether the thesis should be rejected. PhD students are well aware of the high stake here and this wash back on how they present their research. The effect of summative assessment is that students cover up their weak points (Biggs & Tang, 2007, p. 164), and hence avoid going into too deep discussions of what they did wrong or could have done better or differently. Considering the high stake for PhD students it seems quite understandable if they hesitate to evaluate their own work critically in the thesis, but rather prepare to answer the critical questions from the assessors at the defence or viva.

Anne Lee has developed a framework on approaches to supervision, based on interviews with supervisors (Anne Lee, 2008). In her analysis and description of the five approaches (functional, enculturation, critical thinking, emancipation and developing a quality relationship (p. 270-71)), critical thinking stands out as the approach pertaining the most to supporting autonomy, and as students become more independent they can critique their own work. Lee sees the critical thinking approach as the core of supervision. Brodin (2016) links creative and critical thinking as interdependent and argues that both are necessary components of 'doctorateness'.

Another take on autonomy is to view self-regulated learning as a step towards autonomy. Boud (2000) and Boud and Soler (2015) suggest that assessment should build student competences beyond the timeframe of a course or study programme in order to prepare graduates for working life and lifelong learning. Boud coins this as sustainable assessment, and self-assessment plays a vital role in building the competence to make informed judgement of one's own learning. Self-assessment can enhance students' self-regulation and reduce dependence on the teacher (Brown & Harris, 2014; Sadler, 2010). For self-assessment to be valid and reliable there is a need to inform the judgement by standards and criteria, and this makes the link to assessment criteria.

Research into assessment criteria at PhD level

Coming back to the quotations in my opening of the article, autonomous PhD students are expected to be able to judge their own work, but independence (or autonomy) is a vague goal for setting a direction in PhD education, as the supervisor stated.

In Gurr's model, the hands-on vs. hands-off approach depends on the PhD student's development from dependent to *competently autonomous*. Gurr defines 'competent autonomous' as the discipline neutral aim of researcher education: "The PhD process must, therefore, produce graduates with competent autonomy who, independently of their supervisor, are cognisant of the norms, expectations and standards within their discipline and are able to assess their own plans and actions to ensure compliance with these" (p. 85). The supervisors are expected to know the norms, expectations and standards within their discipline, but as shown by e.g. Gerholm (1990), the norms, expectations and standards within a discipline are often tacit, especially for experienced supervisors who have internalised the expectations of the discipline and have become experts (Patel, Arocha & Kaufman, 1999). To quote Nicol and Macfarlane-Dick (2006, p. 206): "Most criteria for academic tasks are complex, multidimensional […] and difficult to articulate; they are often 'tacit' and unarticulated in the mind of the teacher". This can make it difficult for supervisors to be transparent and share intensions and directions. The research into assessment shows that feedback to students must be linked to assessment criteria (Black & Wiliam, 2009; Hattie & Timperley, 2007), and supervisors would benefit from clearer and more explicit assessment criteria to steer and underpin their feedback (Krumsvik, Øfstegaard & Jones, 2016). Pam Denicolo (2003) investigated the assessment of PhD theses in UK through a small survey carried out within the social sciences. She found that 'the degree of consensus about the criteria is low' and recommends that 'the students and supervisors should be provided with clear criteria to guide the process' (p. 90). Mullins and Kiley (2002) found that many examiners confidently used their own internalized criteria when assessing a thesis, often without consulting institutional guidelines.

Research that aims to explicate assessment criteria is an under-researched field, but especially two major contributions are worth mentioning. In her book, 'Making the Implicit Explicit' Lovitts (2007) provides general discipline neutral criteria as well as more specific criteria within a number of disciplines, based on interviews with supervisors in the United States. The other major contribution is a number of articles by Alyson Holbrook and her group in University of Newcastle, Australia. They have researched assessment of PhD theses through analysis of examiners' reports in the large scale project Study of Research Training and Impact (SORTI www.newcastle.edu.au/research-and-innovation/centre/sorti/). Basing the research on examiners' reports rather than interviews gives better insights into what examiners actually do rather than what they remember, perceive or intend to do.

An examiner's report is first and foremost a summative assessment; it is a judgement of the candidate's abilities to live up to the standards of the discipline assessed through the thesis. However, if the summative assessment concludes that amendments are required for the award of the PhD degree, the PhD student needs to know what to improve in order to get the thesis accepted, and formative comments in the report aim to direct the candidate towards the items

that need improvement. The assessment then has a mixture of summative and formative purposes. Holbrook et al. (2014) made an extensive analysis of formative feedback from examiners to PhD students on weaknesses and flaws in theses that were assessed 'less favourable recommendation' in Science and in Education disciplines. They found that more formative comments predicted a weaker thesis, and in Science especially comments on data analysis and on methods predicted a weaker thesis.

In another study Holbrook, Bourke and Fairbairn (2015) specifically analysed examiners' references to theory in science and education disciplines. The six categories of summative comments (positive and negative) resulting from their analysis can provide PhD students with criteria they need to attend to, for instance the coverage of the literature review and considerations of strengths and weaknesses of theories. Similarly, her group studied references to the literature review (Holbrook et al., 2007), but in this study across the full range of disciplines data was not disaggregated according to disciplines. The categories of comments are therefore discipline neutral, and it could be interesting to investigate possible disciplinary differences in reference to categories like coverage, inaccuracy and application, and subcategories of for example coherent use, critical appraisal and connection with own research.

A MODEL OF DOCTORAL SUPERVISION AS FORMATIVE ASSESSMENT

The idea of formative assessment as defined by Harlen (2013) and Dolin et al. (2017) is to involve students in assessing their own work. The distinction between formative assessment and formative feedback in this model is that formative feedback is only a part of formative assessment; it is the feedback that the supervisor (or others) provides to the student with the aim of enhancing student learning. Formative assessment involves the whole circular process of students' activities, evidence of their achievements, judgement and next steps in the learning process, with students involved in interpretation and judgement, and in deciding what and how the next steps should be taken. However, there is a need to translate the model developed by Harlen (2013) and Dolin et al. (2017) to the context of doctoral supervision to increase relevance for PhD students and supervisors.

Translation into the realm of PhD education means that the goals and the learning processes are more fluid, since the research process by nature cannot be foreseen and planned in detail. It is through the research project that competences are acquired and knowledge produced. The overall goals encompass the implementation of a research project with production of new (original) knowledge, and the communication of this in the doctoral thesis. In undergraduate studies students learn (or acquire) already established knowledge, by making sense of it and constructing their own understanding. As Bowden and Marton (1998) argue, research is also learning in the sense that the scholarly community learns new things about the World, and here the PhD student and the supervisor are sometimes on equal ground in that they both learn. In PhD education the learning process cannot be planned the same way, with clear disciplinary assessment criteria and progression steps, but the learning process is steered by the research questions and plans need to be changed as research progresses.

In formative assessment the relationship between the PhD student and the supervisor is essential. While formative feedback is typically provided by the supervisor as the knowing authority (Dysthe, 2002) or critical friend (Deuchar, 2008), or even the examiners who provide comments to the thesis (Holbrook et al., 2014), formative assessment demands that the supervisor is an ally; the PhD student and supervisor are collaborating in a partnership although one is more experienced than the other (Dysthe, 2002). The PhD student – supervisor relationship is also important for formative feedback to work optimally as pointed out by van Rensburg and Danaher (2009), but in formative assessment it is fundamental.

The PhD student is in the centre to emphasise the student centeredness of supervision (c.f. Gurr, 2001) and to indicate that the student is in charge of their own learning process. In PhD supervision this is a complex construction that can be difficult to maintain, since both sides have expectations to the role of the PhD student. Especially students coming to a Northern European university from educational systems where they are expected to 'listen and obey' would indirectly position the supervisor as an authority. Students coming from educational systems characterised by a performance culture with numerous tests may tend to look for judgement rather than feedback to enhance learning (Dolin et al., 2017; Midgley, Kaplan & Middleton, 2001). The construction of roles and responsibilities in a supervisory relationship is a fluid, two-way process (Davies & Harré, 1990; Kobayashi, Grout & Rump, 2015), and this calls for an alignment of expectations (Kiley, 2009).

The model is depicted in Figure 2. In practice the process would not be as formal as depicted, but as a model it can help keeping an overview of the process at a meta-level. The PhD student performs some activity, (1) in Figure 2, be it writing or practical research as part of PhD studies, to fulfil medium term goals as well as the overall goal of PhD education. Especially mid-term goals can be very different dependent on the discipline, and in health and science disciplines the goals are often perceived as completing a research project and publishing the results. Often the research questions rather than competence goals of the PhD education will guide the learning process.

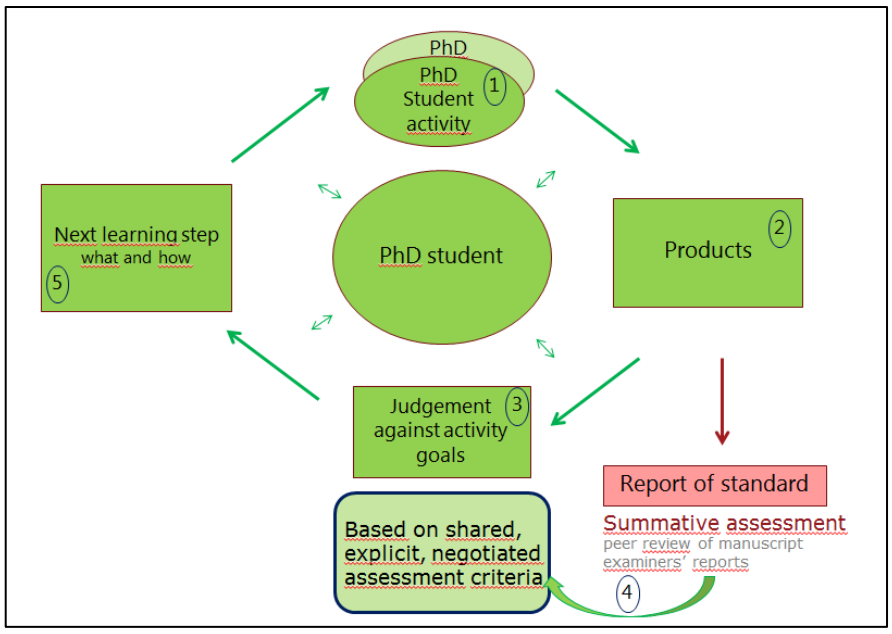


Figure 2. Supervision as formative assessment based on Dolin et al. (2017)

The student activity leads to a product (2). A product is something that can be observed and explicated. If the activity is an experiment in the lab, then the product could be the results, or it could in fact also be the activity itself, observed by the supervisor. Asking questions to probe the student's understanding is yet another way to collect data, however, here it is of uttermost importance to be transparent about the intensions; if students feel evaluated and controlled this may wash back on them and produce students who tend to perform and show their best rather than being open about their doubts and weaknesses, and this will inhibit or lessen the learning process (Tofteskov, 1996). The product has to be observable to enable discussion about fulfilment of the criteria and allow for judgement. In a sense it should provide evidence for judgement.

To reach step (3) the product is judged against criteria pertaining to the relevant goals and the ambitions and capabilities of the PhD student, i.e. both criterion-referenced and student-referenced. This double reference points to a dilemma that many supervisors find themselves in: that they should be both an ally, mentor or coach and guide the PhD student, and also are bound to set the bar high enough, to safeguard standards as gatekeepers of the discipline (Alison. Lee & Green, 2009). The criterion-referenced judgement should use criteria that are explicit, shared and clear, but in PhD education criteria are often internalised and tacit. In formative assessment the verbalisation or even development of the criteria may be part of the process, and point (4) in Figure 2 shows a way to develop or uncover criteria: PhD students in science and health disciplines especially, submit manuscripts to journals and get comments from reviewers. Criteria can be deducted from these reviews, and thereby the criteria are also external to the supervisor and the research group. The same applies to former PhD students' theses that have been reviewed by examiners, as the work by Holbrook and her group indicates. Hence, this is a way to utilise summative assessment for a formative purpose. In situations where the product is research then the criteria are derived from the research questions. While both supervisor and PhD student are in a learning process when producing new knowledge, the supervisor would have greater expertise to assess validity of the research and explicate relevant criteria. The explicit criteria are essential; they are part of the important meta-communication and they convey a direction for PhD students to work more independently.

The use of criteria external to the supervisor and the local research group where the PhD student works ensures that the PhD student is not merely reproducing existing knowledge and methods as in a master-apprenticeship. It ensures the necessary reflectivity and critique of the social practice in the local research environment, and addresses the criticism of the master-apprentice model by for instance Russell (1998).

Notwithstanding the importance of clear and explicit criteria, the process of grasping the nature of quality is complex. A supervisor will often notice particular strengths or weaknesses in a text without referring to specific criteria; as Sadler (2010, p. 546) puts it "drawn from an undefined pool of potential criteria". One way forward can be for the supervisor to argue *why* something is particularly well formulated or *what* and *why* something might be lacking. Another way to induce PhD students into the scientific thinking and the nature of quality is to invite another researcher or the co-supervisor to take part in the scientific discussions. This creates learning opportunities for the PhD student as researchers draw from their more or less

tacit knowledge in discussions with others (Kobayashi et al., 2017). To some extent (some) criteria remain tacit as they are internalised by the PhD student through the participation in scientific discussion.

The judgement of the product is used to decide on the next step in the learning process (5), be it research as learning or development of the PhD student's skills and competences. The formative feedback describes what is needed for moving on in the learning process. The feedback that PhD students get may come from the supervisor, from peers, other colleagues in the scientific community or even from the PhD student herself (Dolin et al., 2017), and indeed the latter is central for self-regulated learning, meta-cognition and autonomy. This is why it is a good idea to ask the student for his/her own assessment as suggested by Handal and Lauvås (2005).

The small two-headed arrows in the model indicate the double role of the PhD student as both provider of products, assessment and feedback and as receiver of the same. The process may not follow the cycle fully in practice. It will often be the case that goals are revisited in the formulation of criteria, and the model should not be seen as stages in a learning process.

When the piece of writing is done, the PhD student might compare that with her first draft and realise how much she enhanced her understanding and her grasp of the topic/method/theory, applying the principles of ipsative assessment (Hughes, 2011). This form of assessment can be highly motivating and give PhD students a sense of standing on more solid ground and build research self-efficacy.

FURTHER RESEARCH

This model for formative assessment in PhD supervision aims to support PhD students and supervisors in the supervisory process. It stresses the importance of criteria, but the model as such does not list assessment criteria. Overall discipline neutral assessment criteria are not sufficient to provide formative feedback and there is a need for more research into operational discipline specific criteria, to supplement the current body of research into criteria at PhD level. There is also a need for research into criteria used in other local contexts to juxtapose and investigate differences arising from different educational systems and goals with PhD education.

Further research will aim is to investigate which criteria examiners use when assessing PhD theses in the different specific disciplines within science in the University of Copenhagen, Denmark. Assessment criteria that examiners are required to use are very general, and may not reflect the specific criteria used in practice in specific disciplines, or may be weighed differently in different disciplines. An analysis of examiners reports can reveal the criteria and the weighing of criteria that examiners tacitly employ. It is intended to use the same research methods as earlier research conducted in University of Newcastle, Australia to allow juxtaposing results. Such research would also enable further research into progression in researcher education and potential disciplinary differences.

ACKNOWLEDGEMENT

I wish to thank Professor Jens Dolin for his inspiration and feedback on my teaching and research, leading to the development of the model targeted at PhD education and supervision.

REFERENCES

Baltzersen, R. K. (2013). The Importance of Metacommunication in Supervision Processes in Higher Education. *International Journal of Higher Education, 2*(2), p128.

Biggs, J., & Tang, C. (2007). *Teaching for Quality Learning at University* (3rd Edition ed.). Maidenhead, UK: The Society for Research into Higher Education & Open University Press.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21*(1), 5-31.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167.

Boud, D., & Soler, R. (2015). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, 1-14.

Bowden, J., & Marton, F. (1998). *The university of learning. Beyond quality and competence*. London and New York: RoutledgeFalmer.

Brodin, E. M. (2016). Critical and creative thinking nexus: learning experiences of doctoral students. *Studies in Higher Education, 41*(6), 971-989.

Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research, 2*(1), 22-30.

Davies, B., & Harré, R. (1990). Positioning: The Discursive Production of Selves. *Journal for the Theory of Social Behaviour, 20*(1), 43-63.

Delamont, S., Parry, O., & Atkinson, P. (1998). Creating a delicate balance: the doctoral supervisor's dilemmas. *Teaching in Higher Education, 3*(2), 157-172.

Denicolo, P. (2003). Assessing the PhD: A constructive view of criteria. *Quality Assurance in Education, 11*(2), 84-91.

Deuchar, R. (2008). Facilitator, director or critical friend?: contradiction and congruence in doctoral supervision styles. *Teaching in Higher Education, 13*(4), 489-500.

Dolin, J., Black, P., Harlen, W., & Tiberghien, A. (2017). Exploring relations between formative and summative assessment. In J. Dolin & R. H. Evans (Eds.), *Transforming assessment – through an interplay between practice, research and policy* (pp. 53-80). Cham, Switzerland: Springer.

Dysthe, O. (2002). Professors as Mediators of Academic Text Cultures: An Interview Study with Advisors and Master's Degree Students in Three Disciplines in a Norwegian University. *Written Communication, 19*(4), 493-544.

Gardner, S. K. (2008). " What's too much and what's too little?": The process of becoming an independent researcher in doctoral education. *The Journal of Higher Education, 79*(3), 326-350.

Gerholm, T. (1990). On tacit knowledge in academia. *European Journal of Education*, 263-271.

Gurr, G. M. (2001). Negotiating the 'rackety brdige' - a dynamic model for aligning supervisory style with research student development. *Higher Education Research and Development, 20*(1).

Handal, G., & Lauvås, P. (2005). Optimal use of feedback in research supervision with master and doctoral students. *Nordic Studies in Education, 2005*(3), 177-189.

Harlen, W. (2013). *Assessment & inquiry-based science education: Issues in policy and practice*: Global Network of Science Academies.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*(1), 81-112.

Holbrook, A., Bourke, S., & Fairbairn, H. (2015). Examiner reference to theory in PhD theses. *Innovations in Education and Teaching International, 52*(1), 75-85.

Holbrook, A., Bourke, S., Fairbairn, H., & Lovat, T. (2007). Examiner comment on the literature review in Ph. D. theses. *Studies in Higher Education, 32*(3), 337-356.

Holbrook, A., Bourke, S., Fairbairn, H., & Lovat, T. (2014). The focus and substance of formative comment provided by PhD examiners. *Studies in Higher Education, 39*(6), 983-1000.

Holbrook, A., Bourke, S., Lovat, T., & Dally, K. (2004). Investigating PhD thesis examination reports. *International Journal of Educational Research, 41*(2), 98-120.

Hughes, G. (2011). Towards a personal best: a case for introducing ipsative assessment in higher education. *Studies in Higher Education, 36*(3), 353-367.

Kam, H. B. (1997). Style and quality in research supervision: the supervisor dependency factor. *Higher education, 34*(1), 81-103.

Kiley, M. (2009). Isn't research just research? What are candidates and supervisors thinking? In A. Brew & L. Lucas (Eds.), *Academic Research and Researchers* (pp. 161). The Society for Research into Higher Education: Open University Press.

Kobayashi, S., Berge, M., Grout, B. W. W., & Rump, C. Ø. (2017). Experiencing variation: learning opportunities in doctoral supervision. *Instructional Science*.

Kobayashi, S., Grout, B. W., & Rump, C. Ø. (2015). Opportunities to learn scientific thinking in joint doctoral supervision. *Innovations in Education and Teaching International, 52*(1), 41-51.

Krumsvik, R. J., Øfstegaard, M., & Jones, L. Ø. (2016). Retningslinjer og vurderingskriterier for den artikkelbaserte ph. d-avhandlingen.

Lee, A. (2008). How are doctoral students supervised? Concepts of doctoral research supervision. *Studies in Higher Education, 33*(3), 267-281.

Lee, A., & Green, B. (2009). Supervision as metaphor. *Studies in Higher Education, 34*(6), 615-630.

Lovitts, B. E. (2005). Being a good course-taker is not enough: a theoretical perspective on the transition to independent research. *Studies in Higher Education, 30*(2), 137 - 154.

Lovitts, B. E. (2007). *Making the implicit explicit: Creating performance expectations for the dissertation*: Stylus Publishing, LLC.

Lovitts, B. E. (2008). The transition to independent research: who makes it, who doesn't, and why. *The Journal of Higher Education, 79*(3), 296-325.

Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology, 93*(1), 77-86.

Molly, A., & Kobayashi, S. (2014). Coachende vejledning – en dynamisk vejledningsstil [Supervision by coaching - a dynamic style of supervision]. *Dansk Universitetspædagogisk Tidsskrift (Danish Journal of Teaching and Learning in Higher Education), 9*(16), 84-95.

Mullins, G., & Kiley, M. (2002). 'It's a PhD, not a Nobel Prize': how experienced examiners assess research theses. *Studies in Higher Education, 27*(4), 369-386.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

Overall, N. C., Deane, K. L., & Peterson, E. R. (2011). Promoting doctoral students' research self-efficacy: combining academic guidance with autonomy support. *Higher Education Research & Development, 30*(6), 791-805.

Patel, V. L., Arocha, J. F., & Kaufman, D. R. (1999). Expertise and tacit knowledge in medicine. In R. J. Sternberg & J. A. Horvath (Eds.), *Tacit knowledge in professional practice: Researcher and practitioner perspectives* (pp. 75-99). London: Lawrence Erlbaum Associates.

Russell, D. R. (1998, 1-4 April 1998). *The limits of the apprenticeship models in WAC/WID research.* Paper presented at the Conference on College Composition and Communication Chicago

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535-550.

Strong, T., Sutherland, O., Couture, S., Godard, G., & Hope, T. (2008). Karl Tomm's Collaborative Approaches to Counselling. *Canadian Journal of Counselling and Psychotherapy/Revue canadienne de counseling et de psychothérapie, 42*(3).

Tinkler, P., & Jackson, C. (2004). *The doctoral examination process: A handbook for students, examiners and supervisors: A Handbook for Students, Examiners and Supervisors*: McGraw-Hill Education (UK).

Tofteskov, J. (1996). *Projektvejledning – og organisering af projektarbejde*: Forlaget Samfundslitteratur.

van Rensburg, H., & Danaher, P. A. (2009). *Facilitating formative feedback: an undervalued dimension of assessing doctoral students' learning.* Paper presented at the Proceedings of the ATN Assessment Conference 2009: Assessment in Different Dimensions: A Conference on Teaching and Learning in Tertiary Education.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

PISA SCIENCE ITEM DIFFICULTIES ACCORDING TO SOCIO-ECONOMIC-CULTURAL LEVEL

Mylène Duclos[1], Florence Le Hebel[2], Ira Noveck[4], Pascale Montpied[2], Andrée Tiberghien[2], Valérie Fontanieu[4], Ira Noveck[5], Jean-Baptiste Van der Henst[5] and Jacques Jayez[6]

[1]UMS LLE ENS Lyon, UMR ICAR, Lyon, France; [2]University of Lyon 1, UMR ICAR, CNRS, LLE, ENS Lyon, France; [3]UMR ICAR, CNRS, Lyon, France; [4]ENS Lyon, IFE, France; [5]CNRS, UMR 5304 ISC, Bron, France ; [6]ENS Lyon, UMR 5304 CNRS, Bron, France

PISA assesses to what degree 15-year-old students have acquired knowledge and skills that are essential for life in society. French results from PISA science 2015 show that the influence of students' socio-economic-cultural status (ESCS) on their performance is one of the highest among OECD countries. The aim of our study is to identify some of the main characteristics of PISA items, which make them difficult to understand for the students according to their ESCS. Our approach combines a quantitative and qualitative analysis. We focus on the performance gap between French students with high and low ESCS. The ESCS index was divided into quartiles (equal groups of 25%). The ESCS 1 group refers to the 25% most disadvantaged pupils while the ESCS 4 group corresponds to the 25% most advantaged students. We made a repartition of items according to their difficulty and interquartile gap (ESCS4-ESCS1). The majority of high interquartile gap items corresponds to the medium difficulty items. In order to find the characteristics that may influence students' scores, we conduct an a priori analysis of the items which we validate by a statistical study on the scores. Several items' characteristics identified in our analysis appear to statistically favor high ESCS students compared to low ESCS students.

Keywords: PISA science 2015, socio-economic level, context.

OBJECTIVES

The aim of the study is to identify some of the main characteristics of science tasks such as those presented in PISA, which discriminate among the students' performances based on their economic, social and cultural status (ESCS). PISA assesses to what degree 15-year-old students have acquired knowledge and skills that are essential for life in society (OECD, 2016). In 2015, the major domain (i.e. the field with the most questions) was scientific literacy. PISA also assesses the students' ESCS index. French results from PISA science 2015 show that the influence of students' ESCS on their performance is one of the highest among OECD countries.

In order to find the characteristics that may influence students' scores, we conduct an *a priori* analysis of the items that we then validate by a statistical study on the scores. This study aims to understand the explanatory power of these characteristics. In particular, we focus on the performance gap between high and low ESCS students. This leads us to hypothesize about the difficulties encountered by low ESCS students in solving PISA Science tasks.

THEORETICAL FRAMEWORK

The first part of the framework is focused on the PISA science framework and the second on

how students, based on their ESCS, understand the different components of the tasks.

PISA 2015 Science framework

The Program for International Student Assessment (PISA) is an international study of 15-year-olds students that every 3 years assesses the knowledge and skills essential for full participation in society (OECD, 2016). In 2015, the major domain evaluated (i.e. the field with the most questions) was scientific literacy. French students' level of scientific literacy has not improved since 2006 and remains within the OECD countries average.

PISA 2015 Science was based on the following components (OECD, 2016):

- Two scientific contexts of the situations on which the questions are based. Context 1 is related to health and disease; natural resources; environmental quality; hazards; frontiers of science and technology; context 2 is related to the self, family and peer groups (personal), to the community (local/national), and to life across the world (global);
- The scientific competencies (Explaining phenomena scientifically; Evaluating and designing scientific enquiry; Interpreting data and evidence scientifically);
- The domains of scientific knowledge. The PISA Science 2015 framework distinguished between "knowledge of science content" (scientific concepts in the domains of Physical systems; Living systems; Earth and Space systems), "epistemic knowledge" referring to an understanding of the role of specific constructs and defining features essential to the process of knowledge-building in science (Dushl 2007) and "procedural knowledge" (knowledge of the practices and concepts on which empirical studies are based)

The items have different formats: simple multiple choice, complex multiple choice or open responses. In parallel to the PISA science test, students completed a questionnaire measuring their Economic, Social and Cultural Status (ESCS) based on three indices: parents' highest occupational status, parents' highest education level in years of education, and home possessions (see Keskpaik & Rocher, 2011).

Students understanding of the tasks' components

Concerning the studies of student understanding, we focused on those related to student understanding of scientific texts. Marin et al. (2007) state that scientific texts are often more difficult for students than narrative ones for several reasons. The lexicon is specialized and they provide insufficient context to clarify the meaning of these words. In the case of PISA units, this comprehension is often crucial for the student to be able to answer the question. These researchers (Marin et al, 2007) also explain that the inferences are essential for the comprehension of a scientific text. Some students would not be able to do this inferential work and some help would be beneficial.

This type of difficulty can make these texts discriminatory for economic-social-cultural status. The results of our previous studies confirm and explore the difficulties faced by low achievers. They show that, when they elaborate and arrive at their answer, low achievers mostly do not

construct correct and stable representations of what the goals of a PISA item are (Le Hebel, 2014; 2016; 2017). Consequently, low-achievers often transform the question in order to be able to answer it. In contrast to high achievers, they are not able to identify the steps they must make between the initial information and the expected final aim of the task. They are not aware of what they are expected to supply - new knowledge for instance – in order to solve the task.

Frequently, PISA science items include somewhat lengthy texts and possibly illustrations that play a crucial role in answering strategies. Therefore, we refer to Delarue-Breton & Bautier (2015) who examined reading literacy and for example showed that low achievers focus on specific elements that echo their experiences or opinions whereas high achievers construct general and generic meaning.

Researchers have already tried to explain what could make PISA science tasks difficult, and also scientific statements in general. Solano-Flores et al. (2015) showed that the characteristics of PISA illustrations could have an influence on student performance. Le Hebel et al. (accepted) showed that the difficulty of PISA Science 2015 items do not necessarily have a high cognitive complexity according to the DOK scale (Webb, 2007). Therefore, other factors are a source of difficulty, especially for students from disadvantaged backgrounds, such as the unfamiliarity of certain components of the context. Indeed, Ahmed et al. (2007) explain that the context of a question can add extra demands. Contextualization of items can also trap students in the sense that they tend to use everyday knowledge (Anahi Da Silva, 2004) rather than scientific knowledge when context places the question in everyday and familiar situations.

METHODOLOGY

To answer our first research question, we need statistical analyses of PISA 2015 data. At the French level, the DEPP (Directorate of Evaluation, Foresight and Performance - French Ministry of Education) can proceed to various statistical analyses. They provided all the information on PISA primary results needed for the present secondary analyses. In PISA, the Economic, Social and Cultural Status (ESCS) provides a measure of the socio-economic status of 15-year-olds, with different component indices (Keskpaik & Rocher, 2011). ESCS was divided into quartiles (ESCS 1/2/3/4). The ESCS 1 group refers to the 25% most disadvantaged pupils while the ESCS 4 group corresponds to the 25% most advantaged students. The DEPP provided us with item success rates according to these ESCS groups. We calculated the difference between the scores obtained by the first and last quartile for each PISA item. The results of this operation showed that the highest interquartile performance gap was 42 points and the lowest was 3 points. We classified the different gaps obtained into four categories as follows with the medium gap group divided into two equal parts:

-low gap (between 3 and 18 points),

-medium gap with low trend (between 19 and 24 points),

-medium gap with high trend (between 25 and 29 points),

-high gap (between 26 and 42 points).

Each PISA item is associated with a proficiency level. These levels are defined *a posteriori* by PISA, that is, only after scoring students' responses. We have estimated the level of proficiency

as a function of the success rate of the items. So, the low difficulty corresponds to a student's highest success rate between 75 and 100%, the medium low trend difficulty refers to a student's success rate comprised between 50 and 75%, the medium high trend difficulty equals a student's success rate located between 25 and 50%, and finally, the high difficulty corresponds to a student's lowest success rate between 0 and 25%. From the four gap categories (obtained on the basis of the calculation: ESCS4-ESCS1) and the four grades of difficulty defined above, we calculated the distribution of items across these eight categories (Table 1).

Our methodology consists of two main steps: an *a priori* items analysis and a statistical study on students' scores.

A priori analysis

First, we took into account the characteristics of six items from PISA 2015 Science framework:

1. Context 1 (Health and disease; Natural resources; Environmental quality; Hazards; Frontiers of Science and technology)

2. Competencies (Explain phenomena scientifically; Evaluate and design scientific enquiry; Interpreting data and evidence scientifically)

3. Knowledge (Knowledge of content of science, Procedural knowledge, Epistemic knowledge)

4. System assessed (for the items assessing knowledge of science content: Physical systems, Living systems, Earth and Space systems)

5. Item format (Simple multiple choice, Complex multiple choice or Open responses)

6. Concerning context 2 (Personal, Local/National, Global) following the *a priori* analysis, we chose to refine this classification as we found PISA coding too general. We defined five components: personal/societal, personal/global, societal/global, societal, global.

Moreover, we add six characteristics defined as follows:

7. The DOK (Depth of Knowledge) corresponding to Webb's DOK levels for science (Webb, 2007). It is a scale of cognitive demands (from 1 to 4) which reflects the cognitive complexity of the question (Le Hebel & al, 2017).

8. Dependence or independence of the question on information available to the students in the item text and/or illustration.

9. "Projection" requirements (or not) meaning that the context of the question prompts the students to project themselves, and conceive the point of view of a community possessing varying degrees of similarity to their own life. Indeed, we observed that some PISA items like the one presented in an upcoming section, require the student to play a role that she is not used to playing at school, for example taking the scientist's place (referring to the researchers' community)

10. If the projection is direct (made explicit in the item text) or indirect (implicit).

11. If the answer is present in the text and/or illustration or not.

12. Text length (word count).

We have analyzed these specific characteristics because we hypothesize that they influence the difference of performance between high and low ESCS students.

In total, we code these twelve characteristics for all 183 items of PISA 2015 Science.

Statistical analysis

First, we observe that the performance gap between ESCS1 and ESCS4 is highly variable (from 0.03 to 0.42) depending on items and that this gap is not linked to item scores (Table 1).

Among the **twelve characteristics** that will potentially explain the different categories of performance gap (on the basis of success rates of ESCS4 group- success rates ESCS1 group), one of them is a quantitative variable (word count), whereas the others are qualitative (ordinal for the DOK or non-ordinal for the item format).

Multiple linear regression models are used to identify an item's characteristics, which influence:

- the students' score (percentage of correct answers for each item in France) of ESCS1 group.
- the students' score (percentage of correct answers for each item in France) of ESCS4 group.
- the difference between high and low ESCS students' scores.

We consider the test allows us to reject the hypothesis of a lack of a characteristic's influence on the scores and on the scores' difference when the p-value is below 0.1.

FINDINGS

First, we present the frequency of each defined category of PISA Science 2015 items by difficulty and by interquartile (ESCS1- ESCS4) (Table 1).

In Table 1, it appears that low and high difficulty items are less discriminative than those of medium difficulty. The medium low difficulty items are almost equitably distributed (between 20% and 29%) and the medium high difficulty tend to be more discriminative (43% for high gap in red circle).

We also calculated differences between each quartile (ESCS2-ESCS1; ESCS3-ESCS2; ESCS4-ESCS3) and it showed different distributions: for instance, in some cases we find a very big difference between ESCS1 and ESCS2 and small difference between ESCS2, ESCS3 and ESCS4, meaning that these items discriminate against the lowest ESCS students (cf. section example of item analysis). On the contrary, some items show a big difference between ESCS3 and ESCS4 and small difference between ESCS1, ESCS2 and ESCS3, meaning that the highest ESCS students perform much better than the other students. The aim of statistical analysis was to connect these qualitative item characteristics presented previously with the quantitative categories described in Table 1 in order to determine the explanatory power of a set of characteristics on the score variations according to the students' ESCS group.

Table 1. Distribution of the PISA Science 2015 items by difficulty and of the gap between ESCS1 and ESCS4.

Distribution of the PISA Science 2015 items	Low gap [3 to 18 points]		Medium low trend gap [19 to 24 points]		Medium high trend gap [25 to 29 points]		High gap [30 to 42 points]		Total
	Fre-quency	relative fre-quency	Fre-quency	relative fre-quency	Fre-quency	relative fre-quency	Fre-quency	relative fre-quency	
Low difficulty [75 to 100%]	9	69%	3	23%	1	8%	/		13 (7%)
Medium low trend difficulty [50 to 75%]	16	20%	23	29%	23	29%	18	23%	80 (44%)
Medium high trend difficulty [25 to 50%]	11	15%	12	17%	18	25%	31	43%	72 (39%)
High difficulty [0 to 25%]	14	78%	4	22%	/		/		18 (10%)
Total	50	27%	42	23%	42	23%	49	27%	183

The data obtained from the national sample of French students participating in PISA 2015 Science are analyzed following our methodology. This treatment results in finding statistical evidence that several characteristics influence performance gaps between the ESCS 4 group and ESCS 1 group (Table 2).

The first column of Table 2 shows the twelve characteristics with their different modalities. For each characteristic, the first modality (first line of each characteristic) is the reference modality for the statistical model. The second column gives the coefficient and the significance for each characteristic modality. When the coefficient is positive, it indicates that the influence of the modality corresponds to an increase in the performance gap between ESCS4 and ESCS1.

A negative coefficient indicates that the influence of the reference modality corresponds to an increase in the performance gap (ESCS4- ESCS1).

Significant results are seen for the characteristics below:

- Dependence on information available to the students in the item text and/or illustration
- Item format
- Types of knowledge
- System (only Earth and Space system)
- Context 1 (only Natural resources)
- Projection (only researchers' community)
- Answer present in item

We were obtained additional evidence from the multiple linear regression models performed on the data from the ESCS1 students' scores and ESCS4 students' scores.

Table 2. Results of Multiple linear regression models (STATA software): Item characteristics by interquartile performance gap (ESCS 4- ESCS 1).

The six characteristics defined by PISA 2015 Science

Item characteristics	Coefficient	Item characteristics	Coefficient
DOK		Knowledge	
1	#	Epistemic knowledge	#
2	-.015	Knowledge of content of science	.056*
3	.019	Procedural knowledge	.052**
4	-.045	System	
Item format		Earth and Space systems	#
Simple multiple choice	#	Physical systems	-.036
Complex multiple choice	.035**	Living systems	-.048**
Open responses	.091***	Context 1	
Competencies		Frontiers of science and technology	#
Explain phenomena scientifically	#	Environmental quality	.027
Evaluate and design scientific enquiry	.008	Natural resources	.031*
Identifying scientific issues	.017	Hazards	.017
		Health and disease	.022

The six characteristics added following the *a priori* analysis

Item characteristics	Coefficient	Item characteristics	Coefficient
Dependence/ Independence		Projection direct or indirect	
Independence	#	No projection	#
Dependence	.041**	Direct projection	.019
Context 2		Indirect projection	0 (omitted)
Global	#	Presence or absence of answer in item	
Global/personal	.034	Absence	#
Societal	.005	Presence	-.073***
Societal/global	.004	Word count	.000
Societal/personal	.007		
Projection			
No projection	#		
Projection “close community”	.021		
Projection “community of know-how”	-.006		
Projection “researchers’ community”	.034*		
Double projection	.019		

*** p-value < 0.01 (difference very significant),
** p-value < 0.05 (difference significant),
* p-value < 0.1 (statistical trend)
# Reference modality

EXAMPLE OF ITEM ANALYSIS

In this section, we will present an example of an item released from PISA 2015 Science “Sustainable fish farming” (Figure 1).

According to PISA characteristics, this item is categorized as following:

-**Competency:** interpreting data and evidence scientifically
-**Knowledge:** knowledge of content of science
-**Context 1:** Environmental quality
-**System assessed:** living systems
-**Item format:** simple multiple choice

And, in the six characteristics added following the *a priori* analysis, this item is defined thus:

-**Context 2:** societal/global
-**The DOK** (Depth of Knowledge): 2
-**Dependence of the question on information available** to the students in the item text and/or illustration.
-**Projection:** knowledge communities and this projection is direct (made explicit in the item text)
-**The answer is present** in the text (framed in orange in the example of item)
-Moreover, the **illustration** of this item is a diagram.

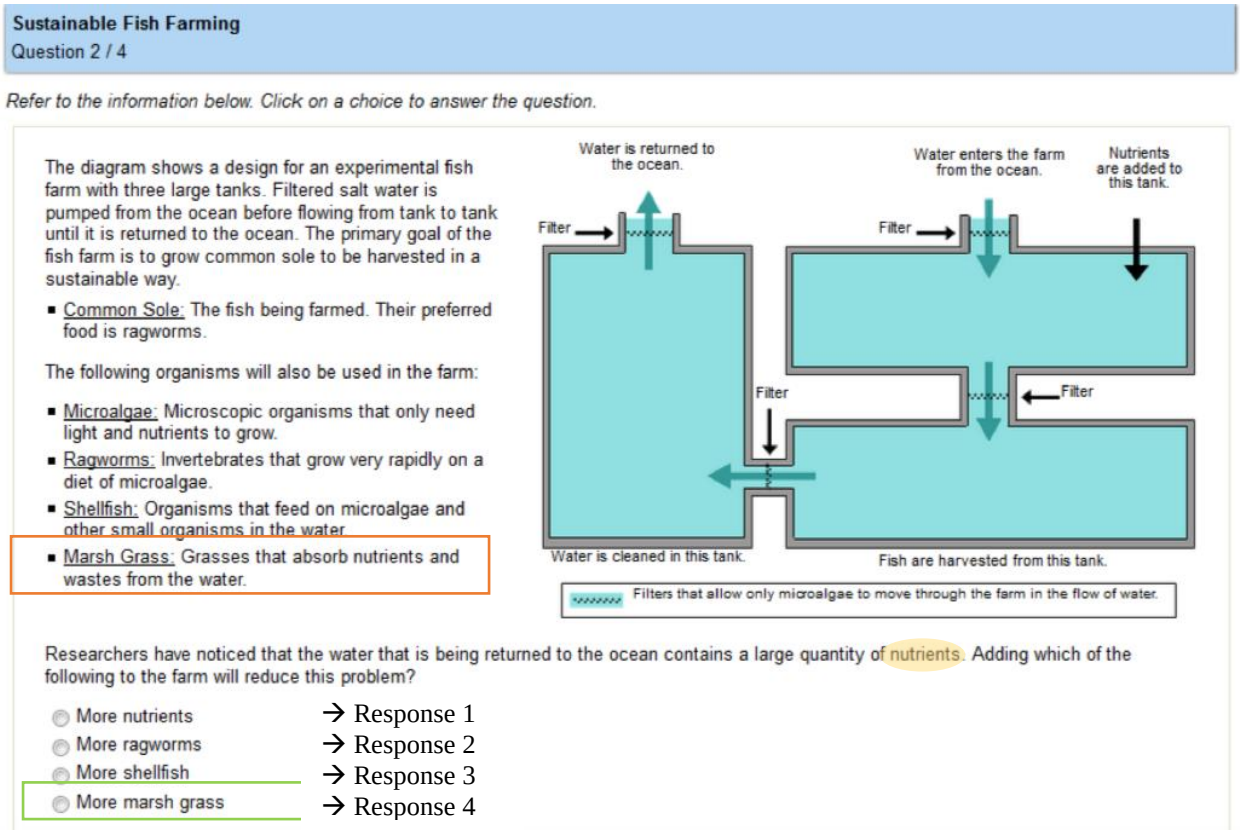


Figure 1. Question 2 of item PISA 2015 Science released "Sustainable fish farming".

This item is located at medium low trend difficulty level (with a success rate of 72%) and with a medium high trend gap (27 points). It belongs therefore to the cluster of 23 items in the red box of Table 1. When we analyzed the results, the ESCS 1 group chose response 1 four times as often as the ESCS 4 group (Figure 2). We propose several potential explanations of the different results.

There is a possible matching for the response 1 (but which influences a wrong answer). Indeed, the word "nutrients" is repeated twice: in the question and response 1 which contains the answer "more nutrients" (circle in yellow in the example of item). Le Hebel et al. (2016) had already observed, on the items of PISA 2006 Science, that low achievers from low ESCS schools use an answering strategy of transforming the question's aim and matching the words from the text of the leading text or the question with the words of the four propositions included in the item.

So, in this "sustainable fish farming" item example, if students did not understand the real aim of the question then they might understand that the sustainable fish farming **lacks nutrients** because the question explains that the water returning to the ocean contains a large quantity of nutrients. So, we suppose that the students may believe that the right response is that the sustainable fish farming will need more nutrients.

Moreover, response 2 is chosen by the ESCS 1 approximately three times as often as the ESCS 4 group and to a lesser extent, the ESCS 1 group also gives response 3 twice as often as the ESCS 4 group (Figure 2). We suppose also that this answer is given because students have difficulty in focusing their attention on the real aim of the item. This can also be explained by a difficulty to find which elements in the flow of information are relevant to answer the question.

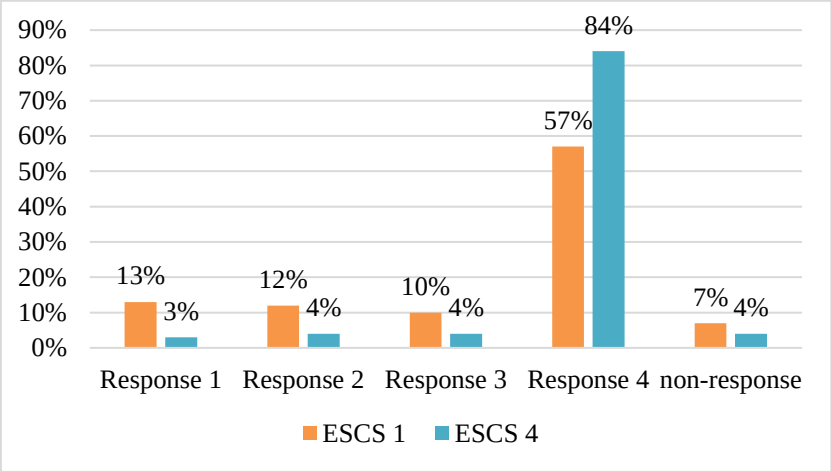


Figure 2. Distribution of responses according to ESCS groups 1 and 4 for the item "Sustainable fish farming"

This item requires a "researchers' community" projection (Table 2) because the question refers to "researchers" and this projection is direct because it is explicitly given in this item. So this context with a direct projection implies the ability to put themselves in the place of scientists. The big gaps in the response choices between ESCS 1 and ESCS 4 could be related to this necessity for the student to make this projection as a scientist and take a scientist's decisions. We had supposed that this projection might potentially be socio-culturally discriminative and our statistical results shows a statistically significant trend Table 2).

The correct answer to the item is response 4 (framed in green). The answer is present in the item because the item gives the exact definition of marsh grasses (framed in orange) and indicates that it is these plants that absorb nutrients. The microalgae also need nutrients but this response choice is not given in the multiple choice, so the question indicates implicitly where to find the response in the item.

As we saw in the "findings" section, when calculating differences between each quartile, we find in some cases a very big difference between ESCS1 and ESCS2 and small difference between ESCS2, ESCS3 and ESCS4, meaning that these items discriminate against the lowest ESCS students.

The example item presented above shows a curve of success rates by ESCS group and we observe that ESCS 1 group literally stands out from the other three ESCS groups with a success rate which differs already very significantly from the ESCS 2 group. Indeed, the ESCS group's success rate is of 57% versus respectively 70%, 76%, and 85% for the three other ESCS groups (Figure 3).

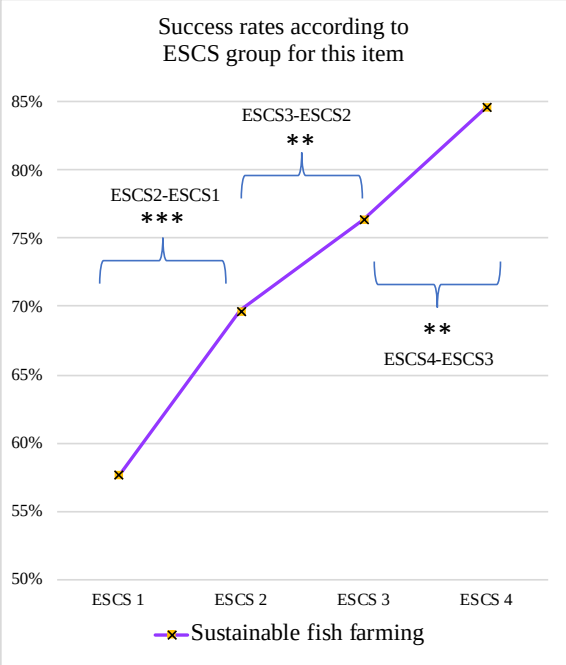


Figure 3. Curve of success rates by ESCS group for the item "Sustainable fish farming"
*** p-value < 0.01 (difference very significant), ** p-value < 0.05 (difference significant)

DISCUSSION

The main findings of statistical analysis show that:

- The information available to the students in the item text and/or illustration provides the ESCS4 students with an advantage compared to the ESCS1 students. It can be interpreted by the fact that ESCS1 students have difficulty in understanding the leading text of the task and in building a representation of the global meaning and goal of the item (Le Hebel & al, 2014). This most likely limits their ability to find the necessary information given in the text in order to build an answer like the example item analysis (presented in a section above) shows it.
- Concerning the item format, open responses increase the performance gap between ESCS1 and ESCS4 students compared to complex multiple choice. ESCS4 students perform more favorably to the open responses format items than ESCS1 students. In addition, the complex multiple-choice format increases the performance gap between ESCS1 and ESCS4 compared to simple multiple choice and favors ESCS4 students.
- Science content knowledge compared to epistemic knowledge gives an advantage to ESCS4 students in comparison to ESCS1 students. However, this result has to be nuanced (indicating a statistical trend). Moreover, the procedural knowledge favors

ESCS4 students significantly compared with ESCS1 students regarding epistemic knowledge. This result deserves to be explored further with other research.

- The items relative to the "Earth and Space systems" domain widens the performance gap between ESCS1 and ESCS4, favoring ESCS4 students.
- Items offering the possibility or requiring that the student engages in a projection called "researchers' community" compared to the items offering no possible projection. It provides an advantage to ESCS4 students who score higher than ESCS1 students (statistical trend). The ESCS1 students, when solving PISA science items, seem to have more difficulty in adopting the point of view of a member belonging to researchers' community when it is required. For ESCS1 students, this community may represent an authority whose role it is not possible for them to play. The other types of projections required by other items do not appear discriminative.
- The fact that the answer is not present in the text and/or illustration widens the performance gap between ESCS1 and ESCS4 students and gives an advantage to the ESCS4 students. On the contrary, an item in which the answer is present is generally more successful in the ESCS 1 group as shown for the item presented in this paper which was successfully answered by 57% of the ESCS 1 group.

Concerning the DOK (Depth of Knowledge), there is no statistical significance as this characteristic affects all students (both ESCS1 and ESCS4). The absence of statistical significance for context 2 (personal/societal, personal/global, societal/global, societal, global) might also be explained by difficulties common to all students whatever their ESCS group. Ahmed & Pollitt (2007) showed that context elements can have an influence on the students' performance but our results did not reveal any effect of this characteristic on the performance gap between ESCS groups.

In line with previous studies described above, our results reveal that many characteristics are likely to interfere with the science task comprehension, in particular with ESCS1 students when the real aim of an item is not understood.

CONCLUSION AND PERSPECTIVES

Our study highlights several items' characteristics, which could permit better understanding of the heterogeneity of the students' performance based on ESCS. In particularly, it could afford a better understanding of the difficulties encountered by disadvantaged students beyond PISA science tasks. Thus, it could help teachers to target these difficulties better in their practice and to take them into account with assessment of low-achievers and their scientific literacy development.

Considering the above results and our previous results (Le Hebel & al, 2017) it appears obvious that in addition to PISA characteristics (competence, type of knowledge, depth of knowledge, PISA context, format), more specific characteristics explain PISA item proficiency.

To refine our research ever more, we have proceeded to recode some characteristics in the a priori analysis in order to refine the statistical analysis. Indeed, as indicated in the example

item, a set of characteristics can be a source of difficulty for the students and so to widen the performance gap between students in group ESCS 1 and 4.

We propose that the low ESCS students have more difficulty in adapting to PISA item situations, due to several levels of unfamiliarity for them. This approach should allow us to target the most representative items to work on in a second step of this project, in which we will focus on low and high ESCS students responding to selected PISA items. We plan to audio-videotape them in order to understand their cognitive processes better and to identify what makes the science task difficult for low ESCS students. This work could help to understand better the difficulties encountered by disadvantaged students beyond complex tasks in science and thus help teachers to target these difficulties in their practice better.

REFERENCES

Anahi Da Silva, V. (2004). *Savoirs quotidiens et savoirs scientifiques ; l'élève entre deux mondes.* Economica - Anthropos Education Poche

Ahmed, A., & Pollitt, A. (2007): Improving the quality of contextualized questions: an experimental investigation of focus, *Assessment in Education: Principles, Policy & Practice, 14*(2), 201-232

Duschl, R. (2007), "Science education in three-part harmony : Balancing conceptual, epistemic and social learning goals", *Review of Research in Education*, *32*, pp. 268-291, http://dx.doi.org/10.3102/0091732X07309371

Keskpaik, S., & Rocher, T. (2011). La mesure de l'équité dans PISA : pour une décomposition des indices statistiques. *Education et formations, 80*, 69-78.

Le Hebel, F., Montpied, P., & Tiberghien, A. (2014). Which Effective Competencies Do Students Use in PISA Assessment of Scientific Literacy? In *Topics and Trends in Current Science Education* (p. 273-289). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7281-6_17

Le Hebel, F., Montpied, P., & Tiberghien, A. (2016). Which Answering Strategies Do Low Achievers Use to Solve PISA Science Items? In *Insights from Research in Science Teaching and Learning* (p. 237-252). Springer, Cham. https://doi.org/10.1007/978-3-319-20074-3_16

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: example of PISA science items, *International Journal of Science Education, 39*(4), 468-487.

Marin, B., Crinon, J., Legros, D., & Avel, P. (2007). Lire un texte documentaire scientifique : quels obstacles, quelles aides à la compréhension ? *Revue française de pédagogie, 160*, 119–131. https://doi.org/10.4000/rfp.786

OECD (2016), PISA 2015 Results (Volume I): Excellence and Equity in Education, PISA, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264266490-en

OECD (2016), PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy, PISA, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264255425-en

Solano-Flores, G., Wang, & Shade, C. (2015) International Semiotics: Item Difficulty and the Complexity of Science Item Illustrations in the PISA-2009 International Test Comparison. *International Journal of Testing*,1 - 15.

Webb, N.L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

SMART: SYSTEMS MAPPING ANALYSIS RESEARCH TOOL

Erica Jablonski[1], Eleanor Abrams[2], Sameer Honwad[1], Elaine Marhefka[1], Robert Eckert[1] and Michael Middleton[3]

[1]University of New Hampshire, Durham, United States of America
[2]University of Massachusetts-Lowell, Lowell, United States of America
[3]Hunter College, City University of New York, New York, United States of America

This paper guides readers through the development, design, and use of a systems mapping-based research tool. The tool specifically helps researchers analyze student understanding of systems thinking using systems maps created for student research on community-based sustainability practices. Initial attempts to analyze student systems maps applied a Structure-Behavior-Function (SBF) approach to quantify essential components of systems represented on student maps, as well as the type of connections between these components (Honwad, et al, 2010). This phase (Phase 1) only partly captured the quality, logic and complexity of connections observed in student systems maps. Consequently, researchers adapted Interaction Analysis (Jordan & Henderson, 1995) for a qualitative approach (Phase 2). Coders found that Phases 1 and 2 were often aligned in the amount of connections and/or student narratives. The qualitative approach better reflected student gains in understanding however because it incorporated the clarity and logic of what was drawn, enabling more meaningful comparisons about the complexity of system understandings before and after inquiry-based research. Because this trend was the case across multiple practices, grades, and student groupings, we believe that the resulting system mapping research tool has the potential for analyzing changes in students' understanding of systems.

Keywords: assessment, researcher-teacher partnership, systems thinking

INTRODUCTION

Over the last century the science education researcher role has evolved from academic theorist to classroom collaborator (Nisbet, 2005). The methods educational researchers use to investigate student learning however have not advanced at the same pace (Wellington, 2015). To establish teacher and institutional buy-in for more collaborative relationships with teachers in the classroom, there is a growing need for dual purpose tools adaptive for both research and pedagogical uses (Kelly, 2004). Using student artifacts as research data is one way researchers and teachers can partner to develop and design tools that are helpful in teaching as well as assessing learning (Merriam & Tisdell, 2015). Systems mapping is one example of a co-developed student-generated artifact to assess student learning of systems and systems thinking (Abrams et al, 2017).

This paper describes an approach designed to analyze group-created systems maps, to assess students' learning of systems thinking and environmental sustainability. While teachers used systems maps for their own teaching and classroom assessment purposes (Abrams et al, 2017), researchers saw an opportunity to gather deeper insight into how students made sense of systems and environmental sustainability phenomena within their own communities.

Our research tool enabled us to systematically analyze student representations of learning in terms of their systems thinking and conceptual growth before and after researching a classroom-selected community practice (e.g., a railroad, shopping mall, afterschool program

building). We believe that the Systems Mapping Analysis Research Tool (SMART) created to understand students' systems maps, has the potential to be adapted for analysis of systems maps across topics and possibly across subject areas where systems thinking is critical to understanding scientific concepts (e.g. engineering processes, systems of the body).

The Methodological Importance and Basics of Systems Map Analysis

The National Science Standards in the United States identify systems thinking and modeling as important concepts that cut across disciplines and grade levels (NGSS, 2013). To facilitate learning about systems thinking, the pedagogical technique of systems mapping has been employed in a range of classroom settings varying in content and academic level from middle school to graduate education (Waters Foundation; Sterman, 1994; Sweeney & Sterman, 2007; Plate & Monroe 2014). However we did not locate prior literature in which systems maps were systematically analyzed using a research tool that focused on assessing students' understandings of different system content across middle school grades.

Systems maps provide students with a way to graphically display their understandings of systems parts and relationships to exhibit how a system operates.A system is all the parts and their dynamic relationships to one another, composing a complex whole. What is considered a system depends on its functional boundaries. Some of the essential aspects of a system include: Components or the different parts (the *'who' and 'what*') that are involved in the function of a selected practice; Connections between components that exhibit how students believe components are related (e.g., inputs and outputs); Feedback loops, visible when outputs are fed back into a system as inputs. The emphasis on different types of interactions helps distinguish systems maps from mind maps which are collections of brainstormed terms related to a concept, or concept maps that are hierarchical constructions of terms related to a concept.

METHODS

For the Supporting and Promoting Indigenous and Rural Adolescents' Learning of Science (SPIRALS) project, systems mapping was used as a pedagogical tool and research artifact. Pre- and post-investigation systems maps were designed as a part of the SPIRALS curriculum (www.spirals.unh.edu) to help students organize and reflect upon what they knew (pre-exploration) or had learned (post-exploration) about the level of sustainability in a selected community practice. In SPIRALS, middle school classrooms select and investigate how a practice in their community may be sustainable. One of the main curriculum goals was to make classroom science relevant to students' everyday lives. The systems map approach was designed in partnership with middle school teachers in New Hampshire after they stated that students needed an activity to help organize their thinking before and after an investigation. A systems thinking approach was determined to be critical for students to understand how different components in a community-based practice were interdependent on each other, and encouraged them to think about community-based practices in terms of relationships and connectedness.

Students worked in groups from two students to entire classrooms to create an initial systems map representing how they believed their community-based practice worked. This map served as a springboard into a scientific inquiry about the selected practice they investigated. In the

"pre" and "post"-exploration map below, 5th grade students focused on whether and how a scenic railway practice functioned sustainably within their community. Between maps students conducted research, in this case by going on a field trip, and communicating with a community practice expert (the railroad operator). In the "pre" map (Figure 1) a few system components are drawn (e.g., rails and "the view"), some of which are connected (e.g., "Coal & Fuel" to "Money") although not always clearly.

The "post" map (Figure 2) displays additional elements and complexity. For example, not only are there more "Materials" listed, but in this map, the materials are integrated into the system. Not only are they inputs into the train but their origins are acknowledged, as with wood resulting from logging. Also added are student narratives about the cost of diesel and the source of fuel (mines in Pennsylvania and West Virginia). Comparing the two maps, it is also apparent that the functional boundary of the Conway Scenic Railroad system, defined by the reach of the systems connections, has expanded. In this post-exploration map therefore students show greater understanding of the parts in the railroad system, as well as their interactions and reach.

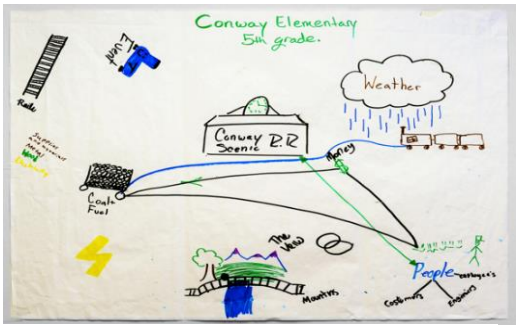


Figure 1. Pre-investigation map

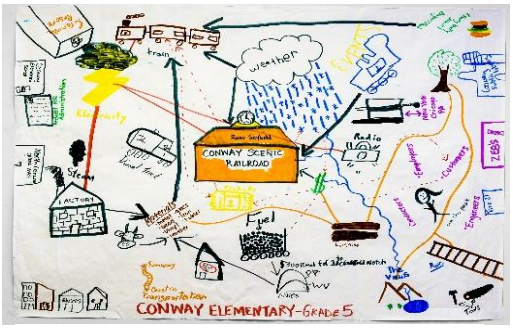


Figure 2. Post-investigation map

Development of SMART

Initial efforts to analyze student systems maps began with an attempt to adapt the Structure-Behavior-Function (SBF) approach to quantify essential elements of systems represented in students maps, as well as the type of connections between these elements (Honwad, et al, 2010). Our coding spreadsheet (Tables 1 & 2) contained essential systems elements we sought to capture. Individual coders completed a spreadsheet for each systems map, listing each system component students named or drew. Coders next identified the number of other components each was connected to, and whether or not they were connected in a way that conveyed a sequential relationship that distinguished inputs from outputs (by directional arrows), or not (by line segments). Essential systems elements were then tallied to identify potential changes between pre and post maps for each student map group. As coders observed inclusion of narratives that conveyed details about components and/or their interactions, a student narrative category was added to the spreadsheet to determine whether changes in the amount of student narratives might also reflect differences in student understandings.

Our attempt to adapt SBF for systems map analyses was not entirely successful for the following reasons:

1 – The community-based practices of SPIRALS systems maps were varied in nature; therefore, from one practice to another, aspects of a system, from components to boundaries,

were not static. SBF has consistently been applied with middle school student understandings of aquatic systems (Hmelo, Marathe & Liu, 2008; Goel, et al 2010; Assaraf & Orion 2010) but not to a broad range of systems.

2 - In SPIRALS, students used multiple ways to describe system components and to indicate relationships between components. Thus, even when students investigated the same practice, maps varied because representations of system components and connections was not rigidly constrained by the curriculum. Two student groups investigated a ski mountain but one emphasized use of the facility (tourists, ski patrol) described primarily using lists, while the other focused on operational inputs and outputs (food, snow making) linked by lines.

3 – Our SBF adaptation , did not accurately reflect the clarity, logic and complexity of connections observed in student systems maps, perhaps in part as a result of the multiple grades using the curriculum and involved in the research.

As a result of these difficulties, we realized the need to develop an analytic approach that was both adaptive to many topic areas and sufficiently comprehensive to capture varied levels of systems thinking understandings. The project's interdisciplinary team of science education researchers, educational practitioners, education psychologists, and scientists responded to this challenge by developing a more comprehensive analytic approach based on the qualitative Interaction Analysis technique (Jordan & Henderson, 1995) involving individual and then consensual coding of human activities, such as artifacts.

Phase 2 was less about determining the number of connections and more about evaluating the nature and quality of connections to make a summative statement about the overall level of systems thinking demonstrated. For phase 2 systems map analysis, individual coders wrote an overall description of a systems map, commenting specifically on the components, overall connections, directionality of the connections, and clarity of the functional boundaries around the system of interest (Tables 1 & 2). After assessing each essential systems element, coders determined an overall systems thinking rating for each map to enable pre-post comparisons. To ensure consistency across coding. each overall systems thinking assessment level (low, low-medium, medium, medium-high, and high) was operationalized to include considerations of map clarity, logic, as well as attempted and successful demonstration of complexity in regard to each essential system element. Systems maps rated low were characterized by lack of practice clarity as when components exhibited little interaction or connectivity. Maps assessed as low-medium were limited to basic depictions of systems thinking, such as primarily implied inputs and outputs. Student maps were assessed as achieving a medium level systems thinking when they indicated a moderate level of connectivity (inputs, outputs) between system components, attempting to depict more advanced systems thinking, but which were not fully developed or clear and logical enough to be considered successful. A medium-high level of systems thinking was attained when maps contain additional complexity as conveyed by better clarity and/or logic such that at least one attempted feedback loop was successful. In contrast, maps designated as displaying high levels of systems thinking contained a majority of successful feedback loops and might also link multiple subsystems together. Following individual coding, coders compared descriptions to establish a consensus on each map's demonstration of systems thinking. Through this process, group understandings of our key

concepts were informed by curricular and expert definitions, but also incorporated evolving definitions and criteria shaped by what was demonstrated in student artifacts themselves.

After the more qualitative second phase of coding was completed with a pre- and post-exploration map pair, we compared its coding with the more quantitative initial phase to determine where the two approaches aligned and what might explain differences in coder determinations in cases of misalignment. While we used mixed methods to provide more valid and trustworthy interpretations of student work, use of multiple coders during both phases 1 and 2 of system map analysis was also employed to bolster the strength of our findings. Individual coding during phase 1 was only conducted once 80% or greater inter-rater reliability was established for greater than 20% of the eventual number of maps evaluated.

Both the quantitative and qualitative analyses highlight changes between students' pre and post systems maps, albeit in different ways. The phase 1 coding in Table 1 corresponds to the pre map in Figure 1. In this table you can see aspects of systems thinking that learners used to demonstrate assumptions about how the railroad they were about to research operated and coder efforts to assess those understandings. Table 1 displays a total of 21 components. Student knew for example that rails, money, and at least one train, were relevant to the railroad. While some components, such as rails were isolated, other components were connected to each other in a directional way by arrows, as displayed between "money" and the use of "coal & fuel". We differentiated between these connections and those that did not signify an understanding of how components were interrelated as in comparing how the "coal & fuel" is connected to the train (via line segment), as opposed to the money. This distinction was based on the assumption that directionality would likely signify a higher level of systems thinking than drawing line segment connections that did not attempt to represent sequencing. At the bottom of each scoring sheet we tallied the components and connections by type, as well as collecting information about whether narratives were used.

Table 1. Phase 1 Pre-exploration map

Map ID	Component	# Connections	# Directional Connections	# Non-directional Connections	Student Narrative
1-C-21-E-M5	Rails	0	0	0	
	Unlabeled (train)	2	0	2	
	Coal and Fuel	3	1	2	
	Money	2	1	1	
	…	…	…	…	…
TOTALS	21	14	5	9	0

In Table 2 we see students more evolved sense of systems thinking relative to the railroad system after their investigation in the phase 1 coding. Rails for example are now connected to other parts in the system instead of being isolated. The train and money ("$"), are more integrated into the system, by being more connected to other parts and by being connected in a directional way. We see in narratives with "fuel" that students have added its source as coal mines, provided information about where those mines are, and detailed fuel costs are per trip.

Table 2. Phase 1, Post-exploration map

Map ID	Component	# Connections	# Directional Connections	# Non-directional Connections	Student Narrative
1-C-21-E-M7	Rails	1	0	1	
	Unlabeled (dollar sign)	7	3	4	
	Train	6	5	1	
	Fuel	2	1	1	$300 diesel for 1 to Crawford Notch
	Mines	2	2	0	PA, WV
	…	…	…	…	…
TOTALS	50	64	31	33	6

Phase 1 alone was not deemed sufficient as components students drew on systems maps were not always clearly related to the practice investigated. Connections were also not always clearly and logically related to how depicted systems functioned. To better address student attempts at greater system complexity, phase 2 captured a more holistic assessment of the same essential system elements. It also incorporated explicit evaluations of clarity, logic, and complexity, and revealed that narrative versus purely graphic representations can demonstrate leaning.

Phase 2 coding of the same pre and post maps (Figures 1 & 2) demonstrates similarities and differences of this approach compared to phase 1. Coder assessment of the pre map (Table 3) displayed a low-medium level of systems thinking. The phase 2 post map in Table 4 is assessed as demonstrating marked improvement in students' systems thinking. Coders agreed to greater numbers of systems components and connections than in the pre map (Table 3) and also to gains in clarity and logic relative to the railroad practice. The overall systems thinking assessment of medium-high reflects partial success in displaying advanced systems thinking, in sub-systems and linkage chains as well as a moderate level of connectivity between system components that are clear and logical. As there were still a number of unclear, illogical or unintegrated components on the map, it was not assessed as demonstrating a high level of systems thinking. Although the conclusions from phases 1 and 2 may align, comparison of the tables above demonstrates the greater utility of the phase 2 assessment.

RESULTS

As part of the SPIRALS project, pre and post systems maps were collected from 10 participating research sites in rural (9) or indigenous (1) communities between Spring 2015 and Fall 2016. Because the curriculum encouraged site-appropriate adaptations, some maps were created at the classroom level while others were created by small groups of 2 to 6 middle school students in grades 4 through 8. Our analytic sample consists of 19 pairs of matched pre and post maps. To reduce undue influences resulting from the loss or introduction of a new student into a small group, only full class-level maps or those with identical students creating both pre and post-exploration systems maps were included in our analysis.

Table 5 displays how pre-to-post systems map assessments of student understanding aligned and diverged. Our analytic sample contained 15 map pairs (79%) for which systems thinking improvements were observed, three map pairs (16%) for which substantial improvements were not observed, and one map pair (5%) for which demonstration of systems thinking declined.

Table 3. Example of phase 2 Pre-exploration map coding

Map ID	Community Practice Clarity	Components	Overall Connections between Components	Inputs, Outputs and Feedback Loops *(Directional Connections)*	Functional System Boundary	Overall Systems thinking
1-C-21-EM5	There is a ***clear*** indicator that students are exploring a railroad practice in their town. What appears to be a train station is labeled "Conway Scenic RR."	Although most components are clear & logically related to a railroad practice, some such as "weather" are less so.	There seems to be a recognition that resources are interconnected in some way but there are also isolated components that indicate less clarity around the connections between other components.	There are some inputs, outputs & 2 attempted feedback loops. Both appear clear & logically related to this system. Clarity: There may be missing components or detail to help understand the nature of drawn relationships. Complexity: Interactions do not appear highly complex.	Boundary of the map appears to be local scenic railway.	**Low-Medium** level systems thinking. Student attempt to show inputs and outputs, however in a simplistic way.

Table 4. Example of phase 2 post-exploration map coding

Map ID	Community Practice Clarity	Components	Overall Connections between Components	Inputs, Outputs and Feedback Loops *(Directional Connections)*	Functional System Boundary	Overall Systems thinking
1-C-21-EM7	There are more clear indicators that students are exploring the railroad practice as the train station labeled "Conway Scenic Railway" is now a central component in the system.	Significant increase in components, most of which contribute to the community practice. Remaining components help contextualize community. Complexity is improved because of more components, more clarity and more detail (labels & detailed labels) about components.	There is a dramatic increase in complexity of map connections (some non-directional, several unidirectional), and multiple components connected in multiple ways to others. There are some isolated components such as "tools" however, it is clear they could be related to the railway practice. The practice also incorporates many subsystems related to the railroad.	There is an increase in the number of inputs and outputs along with the same number of attempted feedback loops. There are also attempts at relating linkage chains that contain many components among the same theme. Most of the inputs and outputs and feedback loops are logical and make sense.	Boundary is clear and extended including connections to non-local inputs.	Medium-High Although students did not provide narrative to show system functions, labels, illustrations & organization help coders understand attempted connections & feedback loops. Complexity reflected in inter-connectivity between components was high & there were several interconnected linkage chains representing potential subsystem.

Coders' qualitative and quantitative evaluations of systems thinking were aligned for the map pair which declined and in the majority (58%) of map pairs showing improvement. Discussion

of maps pairs for which there was no substantial change will be addressed following interpretation of map pairs for which change was observed.

Table 5. Pre Post-exploration Map Comparison between Coding Phases 1 and 2

Pre- to Post-Exploration Systems Maps Compared	PHASE 1				PHASE 2
	Component Difference	**Connection Difference**	**Directional Connection Difference**	**# of Student Narrative Difference**	**Overall Systems Thinking Difference**
1-A-2&3-D	65	32	31	5	Improved
1-A-4&5-156&178-E	-18	-37	19	-11	Improved
1-A-4&5-157&166&181-E	-20	53	53	6	Improved
1-A-4&5-160&175-E	-22	-18	-10	2	Improved
1-A-4&5-162&174-E	1	22	18	-7	Improved
1-A-4&5-164&179-E	21	38	42	-6	Improved
1-B-10-3&8-D	11	21	0	11	Improved
1-C-21-E-M7	29	50	26	6	Improved
1-I-34-G&H531&538&534	-5	27	1	2	Improved
1-J-38-F-412&413&416	7	74	74	4	Improved
2-AA-45-D&E&F	5	19	19	53	Improved
1-O-49-D&E&F-671, et al.	21	-18	-15	4	No Change
2-BB-55-E	-68	-96	-111	0	Improved
1-R-61-G	-19	-2	0	-3	Improved
1-S-62-E&F-1241, et al.	6	29	29	3	No Change
1-S-62-E&F-1243, et al.	1	4	4	10	Improved
1-S-62-E&F-1247, et al.	-10	-8	-8	9	Improved
1-S-62-E&F-1248, et al.	-4	-4	-16	-12	Declined
1-S-63-G&H-1267, et al.	2	32	32	-16	No Change
Agreement on Systems Elements in both Phases	9	11	11	11	
% Agreement by Element	**0.47**	**0.58**	**0.58**	**0.58**	

Because the overall qualitative assessment tended to align with the quantitative assessment across essential system elements (components, connections, and directional connections) in only about two-thirds of the cases (63%), the coding patterns for map pairs that were not aligned were explored. Both phases of coding for the eight map pairs that were least well aligned (5 showing qualitative improvement in phase 2; and 3 assessing no substantial change between pre and post maps) were examined to explain the rationales behind coding divergences. Doing so revealed that connections were most influential in both coding phases. Focus on the nature of connections between components is logical because the complexity of interactions between system components has previously served to differentiate between lower and higher levels of systems thinking attainment (Honwad, et al, 2010; Assaraf & Orion, 2010).

In most (10) of the 16 instances when change was determined between pre and post maps, the numbers of connections and/or directional connections identified in phase 1 were consistent with the direction of the overall assessment in phase 2. In one case (1-A-156&178-E), although the number of components and connections between pre and post maps declined, the number of directional connections increased. This finding is consistent with the idea that systems thinking beyond the most basic level is determined by the complexity of the relationships that students draw, hence improvements are more likely demonstrated through student attempts to convey *how components are related* and *not just which components* are related.

In two other instances when overall improvement was detected between pre and post maps in phase 2 but was inconsistent with declines in essential systems elements in phase 1, increases were however observed in the amount of student narratives (1-S-62-E&F-1247, et al. and 1-A-4&5-160&175-E). The following excerpts from consensual coder syntheses illustrates how narratives can provide sufficient explanations about the nature of interactions between components to compensate for drawings that failed to do so alone:

"Narrative descriptions of how system functions and clarity of understanding of this system all improved on this map, which enabled identification of 3 in 10 successful feedback loops…use of bidirectional arrows however added some confusion." (1-S-62-E&F-1247, et al.).

"Students include narrative showing improvement in their understanding from the first map… Clarity and logic of both components and connections improved drastically…especially due to elimination of the incorrect use of bi-directional arrows." (1-A-4&5-160&175-E).

These excerpts demonstrate how counts of attempted demonstrations of directionality can be misleading as well as how student use of narratives can help clarify the nature of relationships between components that are not always clear and logical from drawings alone. Although narratives were not always clear, they were rarely illogical and hence tended to improve upon drawings that were ambiguous.

In the two remaining maps whose changes were not aligned between the two coding phases (2-BB-55-E and 1-R-61-G), the qualitative coding identified improvement while the quantitative coding, did not. This difference is seen in phase 1 by the lack of change, in these maps on the quantitative measure of student narrative or amount of directional connections, respectively. Comparison of qualitative pre- and post- map coding is instructive in determining why it may be the more valid measure in this instance. Coders indicated in the pre-research map that, "No clear practice…is evident " and that "the many directional connections…are unclear in how they contribute to an overall system". Meanwhile, the follow-up map is described as, "an improvement…in terms of complexity, specificity…and purposeful directionality between different components in the system in a basic but logical linear way." These summaries demonstrate how the *quality* of interactions between components, that designate higher level systems thinking, was more accurately evaluated by considering the nature of directional connections to contextualize their amount.

The last map pair which were misaligned between the two coding phases (1-R-61-G), reinforces the value of the qualitative assessment. Coders explained on the original map that, "inputs and outputs are missing" and thus there was, "little indication of knowledge about

[system component] interactions." Although in the follow-up map coders indicated that, "Relationships are… [still] non-directional" they added that, "[r]elationships among components [and subsystems] are all inferred". In our coding, directionality was initially expected through the use of arrows, as this is how it was presented in the curriculum. In phase 2 coding however adjustments were made to distinguish between graphic display of directionality that were considered ideal, and inferred directionality, which was viewed as less ideal, when narratives indicated inputs and outputs but student drawings did not. Inclusion of inference based on student narratives enabled coders to acknowledge attempts to exhibit greater complexity in their systems thinking even when they were not capable of fully demonstrating it as intended. In this case it enabled coders to conclude that, "Complexity is inferred within the lists, but is not demonstrated in any relational way."

When no substantial change was detected between pre- and post-exploration maps, improvements in some essential systems elements along with declines in others appeared to play a role. Although two of these three pairs displayed this mix of outcomes in phase 1 (1-O-49-D&E&F-671, et al. and 1-S-63-G&H-1267, et al.), these discrepancies again were better addressed through phase 2 coding. Reviewing the map for which phase 1 and 2 codings were completely opposite shows how this can the case. For this map pair, 1-S-62-E&F-1241, et al. consensual coding resulted in the conclusion that, "Complexity improved" but, "clarity and logic decreased". The mixed result was then attributed to the observation that "Students increased complexity with more attempted feedback loops, as well as an attempt at adding another perhaps interrelated health subsystem related to composting. With these attempts the students at times sacrificed clarity and logic."

These coder comments illustrate our finding that although in some post-exploration maps students attempted to demonstrate greater complexity through use of directional arrows and loops connecting multiple components, these efforts were usually not entirely successful. When narrative was not provided to supplement such drawings, the number of successfully depicted interactions between or amongst components declined. Attempts at greater complexity through subsystems followed a similar pattern. We suspect that the grade level may help explain the need for assessments that incorporate credit for attempts at complexity while acknowledging as well whether or not such attempts succeed or fail.

DISCUSSION AND CONCLUSION

Because the value of cross-cutting concepts such as systems thinking have been codified in educational standards (NGSS, 2013), it is important for researchers to determine the extent to which educational efforts have produced student learning. Although prior research had assessed systems thinking learning quantitatively in 4th and 7th grades (Honwad, et al, 2010; Assaraf & Orion, 2010), our effort to assess systems thinking in this manner ran counter to the impressions of the inter-disciplinary team conducting our coding. As a result, to assess students systems learning for the SPIRALS community-based sustainability curriculum, the research team developed a qualitative coding procedure to run in parallel to our quantitative coding. Although both approaches captured a range of outcomes, coders expressed more confidence in the qualitative approach that was produced in a grounded manner, iteratively from content analysis of student work itself. This process was also preferred as it produced thick and rich

descriptions to substantiate coder evaluations of the essential aspects of systems. These descriptions served as the basis for a more nuanced coding scheme as well as providing evidence to explain the differences between the two phases of coding, and our conclusion that the qualitative approach was more appropriate to our student sample.

Although our initial assessment efforts attempted to adopt relevant prior systems thinking research, our study was different because it was aimed at enabling marginalized rural and indigenous student communities to select local community practices to facilitate student engagement. As a result, unlike prior studies, our research involved analysis of different systems, as well as varied grade levels and settings (i.e., public school, charter school, private school, and an afterschool program). We believe that the broader range of settings and content areas address by our study may explain the necessity to devise a novel method to analyze maps in a consistent way that could be applied to different grades and subject matter. The qualitative assessment did appear successful in conveying a range of outcomes across grade levels and appeared to indicate that the curriculum may have been more successful when implemented in classrooms at one grade level, rather than with mixed grades and likewise may be more successful in public school settings than in charter schools. We did not have sufficient data (1 case each) to draw conclusions about the private or afterschool program participants.

There are other possible explanations for the limitations of the curriculum in helping students achieve better systems thinking outcomes as well as limitations to our assessment tool that must be acknowledged: This tool was created for a particular community-based sustainability curriculum intended for middle school students in rural and indigenous settings. Although systems thinking content experts contributed to the tool's development, the tool evolved as it was observed that many students were unable to attain success despite attempts to display greater complexity. It would be beneficial therefore to test the tool in its current iteration on a new, larger sample of systems maps that again cut across grade levels and content areas to determine its adaptability and address the need for a tool that can be flexibly applied.

ACKNOWLEDGEMENT

This work was supported by the United States National Science Foundation AISL Award #1223703.

REFERENCES

Abrams, E., Middleton, M., Honwad S., Jablonski, E., Koper, M., Eckert, R., Varner, R., & Thelemarck., C (2017). Using systems mapping to plan scientific investigations. *Science Scope, 40*(5), 25-31.

Assaraf, O.B-Z., & Orion, N. (2010). Systems thinking skills at the elementary school level. *Journal of Research in Science Teaching, 47*(5), 540-563.

Goel, A.K., Vattam, S.S., Rugaber, S., Joyner, D., Hmelo-Silver, C. Jordan, R., Honwad, S., Gray S., & Sinha, S. (2010). Functional and causal abstractions of complex systems. Paper presented at the annual conference of Cognitive Science Society, 2010.

Hmelo-Silver C. Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: expert-novice understandings of complex systems. *Journal of Learning Sciences, 16*, 307-331.

Honwad, S., Hmelo-Silver, C., Jordan, R., Eberbach, C., Gray, S., Sinha, S., Goel, A., Vattam, S., Rugaber, S, & Joyner, D. (2010). Connecting the visible to the invisible: Helping middle school

students understand complex ecosystem processes. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, Portland, Oregon. pp. 133-138

Jordan, B., & Henderson, A. (1995). Interaction Analysis: Foundations and Practice. *The Journal of the Learning Sciences*, *4*(1), 39-103.

Kelly, A. (2004). Design research in education: Yes, but is it methodological?. *The Journal of the Learning Sciences*, *13*(1), 115-128.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Nisbet, J. (2005). What is educational research? Changing perspectives through the 20th century. *Research Papers in Education*, *20*(1), 25-44.

Plate, R., & Monroe, M. (Winter, 2014). A structure for assessing systems thinking. *The 2014 Creative Learning Exchange*, *23(1).*

Sharma, P., & Xie, Y. (2008). Student experiences of using weblogs: An exploratory study. *Journal of Asynchronous Learning Networks*, *12*(3), 137-156.

Sterman, J.D. (1994). Learning in and about complex systems. *System Dynamics Review*, *10*, 291-330.

Sweeney, L.B., & Sterman, J.D. (2007). Thinking about Systems: Student and teacher conceptions of natural and social systems. . *System Dynamics Review*, *23*, 285-312.

Systems Thinking. (n.d.). *Waters Foundation.* http://watersfoundation.org/systems-thinking/definitions/

Wellington, J. (2015). *Educational research: Contemporary issues and practical approaches*. Bloomsbury Publishing.

ESERA 2017 Conference
Dublin City University, Dublin, Ireland.
21st - 25th August 2017

PROFESSIONAL QUALITY ASSESSMENT OF THE CROATIAN STATE WRITTEN EXAM IN BIOLOGY

Ines Radanović[1], Žaklin Lukša[2], Valerija Begić[3], Mirela Sertić Perić[1] and Diana Garašić
[1]Faculty of Science, Zagreb, Croatia
[2]High School Josipa Slavenskog, Čakovec, Croatia
[3]Primary school Sesvetski Kraljevec, Zagreb, Croatia

The aim of the tool presented in this study is to enable teachers' qualitative analysis of the questions within the Croatian state written exam in biology, and the eventual corrections of the questions before their application in the student assessment. We have identified the two basic categories that determine the question quality: 1) the importance of questions (regarding the profession, life, curriculum, critical thinking), and 2) the influence of questions (i.e., shape and intelligibility of the questions) on students' answers, logical reasoning and further learning path. The tool we have developed was tested for its effectiveness on a sample exam designed for students aged 13. A correlation between logical reasoning and the "importance-of-questions" categories, and the success rate of the exam was observed. This simple tool has proven to be effective for both teachers' self-assessment and peer evaluation.

Keywords: cognitive levels, question relevance for science literacy, influence of questions to answers

INTRODUCTION

Most classroom teachers prepare and administer a series of (non-)formal (i.e., teacher-made) exams during the school year, which often enclose questions with many construction mistakes, especially essay questions (e.g., Marso & Pigge 1988). Thus, there is a growing need for greater quality control in the design and implementation of the students' performance assessments (Dunbar et al. 2009). A tool for the expert question quality assessment in Croatia (representing a developing country regarding its national practice in advancing science literacy and national curriculum) was for the first time designed for the needs of professional quality assessment of the state biology exams (Radanović et al. 2010). In designing the Croatian tool, the following criteria, recognized as "fruitful areas" to seek the question validity evidence, were considered: question content, internal structure and response process, as well as exam scores' relationship to other variables measuring various students' domains, and overall learning success and achievement (Downing, 2003). From its first use, the Croatian tool has been continuously developed through the application within research, as well as within teaching, i.e., in designing written biology exams (Radanović et al. 2011, Begić et al. 2016, Radanović et al. 2017a,b). Thus, since the launch of the Croatian tool, some assessment elements that should encourage teachers to better prepare exam questions have been introduced, and the question quality has steadily increased. The aim of the tool presented within this paper is to enable teachers' qualitative analysis of the questions within the Croatian state biology written exams, and the eventual correction of the questions before their application in the student assessment. An additional aim is to enable the qualitative question analysis in order to more comprehensively interpret student results within the written exam.

METHOD

Based on years of experience in the usage of the question analysis with the assistance of experienced biology teachers, we have developed a tool for assessing the quality of biology written exam questions. The question quality analysis involved a multiple teacher assessments and a collective final consensus-based assessment (MacCann et al, 2004). Elements and criteria for the expert question quality assessment (Table 1) were determined by three point Likert Scale (Cohen et al. 2007).

By shaping the question assessment categories, we relied on the grounds of the PISA project (OECD. 2015) defining science literacy as the ability to engage with science-related issues, and to use scientific ideas, natural science knowledge and evidence-based conclusions as a reflective citizen (Bellová et al. 2017). We defined the two basic categories determining the quality of questions: 1) the importance of the questions (i.e., elements of science literacy) and 2) the influence of questions on students' response.

The **importance of questions (Qim)** category was specifically linked to the importance of questions for the development of science literacy and basic biological concepts (i.e., students' reasoning and conceptual development). By assessing the elements of this category, a three point scale with value range 'unimportant – moderately important – important'was used (Table 1). The assessment elements within this category were the following:

A - **importance of questions for the profession** (IP), i.e., biology – enquiring how much is the knowledge needed for answering the question important and relevant for the development of basic biological concepts, conceptual development and achievement of biological competencies;

B - **importance of questions for life** (IL) – enquiring how much is the knowledge needed for answering the question important and relevant for basic biological literacy and can a student apply that knowledge in present or future life (context-rich questions);

C - **importance of questions for the curriculum** (IC) – enquiring how much is the knowledge needed for answering the question important and relevant for development of the competences foreseen by the curriculum, and conceptual understanding of the biological terms and concepts built-in the prescribed national curriculum;

D - **importance of questions for critical thinking** (ICT) – enquiring how much is for answering the question important reflective thinking focused not only on understanding certain terms and theories, but also on decision making, reasoning and evaluating certain life facts, attitudes and actions; also serves for the assessment of the students' creativity and application of the natural science methodologies, epistemological knowledge, introspection and evidence-based inference; within questions enquiring reproductive knowledge and literature understanding, as important questions are considered those demanding analysis and/or synthesis of basic biological facts extracted during the initial information sorting.

The second category – **influence of questions on students' answers (Qin)** – was closely linked to the influence of the question form, structure, wording and context on the student answering (Table 1).

Table 1. Elements and criteria for the expert question quality assessment.

<table>
<tr><th>Question quality (QQ)</th><th colspan="2">The importance of questions</th><th colspan="2">The influence of questions on students' answers</th></tr>
<tr><td rowspan="5">1 = BAD
2 = ACCEPTABLE
3 = GOOD</td><td>Assessment elements of the science literacy</td><td>Scale of the question importance</td><td>Assessment elements of the influence of questions</td><td>Scale of the question influence</td></tr>
<tr><td>A - importance of questions for the profession (IP)</td><td rowspan="4">1 = unimportant
2 = moderately important
3 = important</td><td>E - question shape/type (QS)</td><td rowspan="4">1 = strongly influences
2 = moderately influences
3 = weakly influence</td></tr>
<tr><td>B - importance of questions for life (IL)</td><td>F - question intelligibility (QI)</td></tr>
<tr><td>C - importance of questions for the curriculum (IC)</td><td>G - students' logical reasoning (SLR)</td></tr>
<tr><td>D - importance of questions for critical thinking (ICT)</td><td>H - students' further learning path (SLP)</td></tr>
<tr><td>(Qim+Qin)/2</td><td>IMPORTANCE OF THE QUESTION (Qim)</td><td>(A+B+C+D)/4</td><td>INFLUENCE OF THE QUESTION (Qin)</td><td>(E+F+G+H)/4</td></tr>
</table>

The assessment elements within this category were the following:

E - **question shape** (QS) – enquiring technical characteristics of the question (information necessary to solve the task): whether the question contains unnecessary and/or distracting information/figure/scheme irrelevant for answering the question; whether the question (text) length and the relevant supplements are in accordance with the question cognitive level; whether the distractor length within the question is consistent; whether the question avoids or accentuates negations; whether the graphs/figures/schemes attached to the question are clear, accurate and adjusted to student age; whether the question stimulus contains all the necessary information needed for answering the question based on learning outcomes prescribed by the relevant curriculum; whether the question scores are matched with the question requirements.

F - **question intelligibility** (QI) – enquiring the adjustment of the question to students' age and understanding; this element could be additionally checked by questioning the following: is the question imprecise, suggestive, confusing, and/or contains conceptually homogeneous distractors and/or too many technical/expert terms irrelevant for shaping an answer; it should be borne in mind that a higher cognitive level requires highly developed literacy for understanding questions and supplementary material

G - **students' logical reasoning** (SLR) – enquiring whether the students' logical reasoning (without students' understanding of the questioned concept) could affect answering the question;

H - **students' further learning path** (SLP) – enquiring whether the question requires additional learning/experience besides the prescribed curriculum (and/or details irrelevant for conceptual understanding) and how much could it affect the answer; whether the question is focused on facts, which are not emphasized during biology classes and/or are not crucial for conceptual understanding of basic biological concepts, but could be acquired by additional learning/experience (by preparing questions for gifted students, the additional learning paths are acceptable, but only to evaluate the level of upgrade of the basic biological concepts and

their application in solving more complex tasks – not to burden the students by memorizing additional terms).

Besides the question quality, the teachers additionally assessed the questions' weight, using the following scale: 1) easy; 2) moderately hard; 3) hard questions. For each question, the cognitive level was assessed according to Crooks (1988), so the questions were attributed to: 1) reproduction; 2) application of knowledge and conceptual understanding; or 3) problem solving.

By harmonizing the statements using an unambiguous numerical scale (Table 1), it was possible to make a more comprehensive question quality assessment, which was initially unfeasible because of the two adverse scales (Table 2) assessing the importance of questions, and the influence of questions on students' answers separately by averaging the scales' scores.

Table 2. Initial scaling for the question quality assessment.

ASSESSMENT CRITERIA	1	2	3
IMPORTANCE OF QUESTIONS	unimportant	moderately important	important
INFLUENCE OF QUESTIONS ON STUDENTS' ANSWERS	strongly influences	moderately influences	weakly influence
QUESTION QUALITY	**bad**	**acceptable**	**good**

The effectiveness of the developed tool was estimated by 4 teachers by means of 148 biology written exams, each comprising 23 questions (Cronbach's alpha = 0.583, n = 148, SE = 0.048, 95% CI = 0.48 to 0.67) targeted for students aged 13 (Begic et al., 2016). Statistical analysis was done by StatsToDo (Chang, 2014), and correlations are interpreted according to Hopkins (2000).

RESULTS

Out of 148 exams encompassed by this study, 91 were written by girls and 57 by boys. Regarding the gender ratio ($Mf = 63.18 \pm 10.97$; $Mm = 66.46 \pm 10.93$), there were no significant differences in the exam performance (i.e., SSR, student success rate). Students successfully answered 3 to 20 questions of the exam (Figure 1). Most students (16%) successfully answered 12 questions, reaching 67.7% of the total points.

Spearman Rank Order Correlation was proven significant for logical reasoning (SLR) and the importance-of-questions (Qim) categories in relation to the success rate of the exam ($\rho = 0.44$, $p < 0.05$).

Student success rate (SSR) of the written exam used for testing our tool was moderately negatively correlated to cognitive level (CL), indicating that students performed better in answering questions of lower CL. Furthermore, SSR was highly correlated with question shape (QS) and influence of questions on students' answers (Qin), suggesting that better formulated questions yield higher answering rate, having a lesser influence on students' answers.

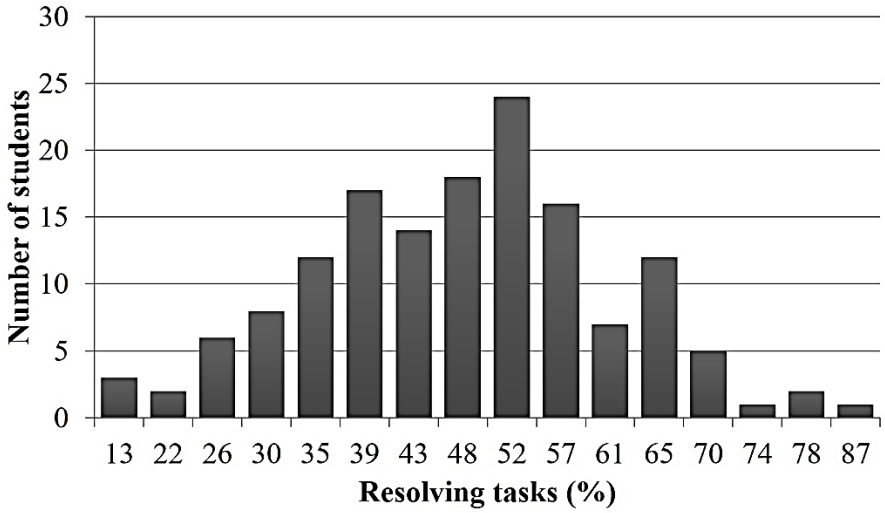


Figure 1. Student success rate of the exam (SSR)

Our results indicate that the questions of lower CL were well-shaped and easily understood by students. For answering these questions, students did not require additional learning/experience besides the prescribed curriculum, and the questions could be successfully answered by applying basic biological concepts. This was additionally corroborated by the moderate correlation between the importance of questions for the profession (biology) (IP) and the importance of questions (Qim) for the development of science literacy as well as by the negative correlation between Qim and question shape (QS) / intelligibility (QI). The observed correlation trends suggest that biologically important questions were less intelligible to students – probably because they were more 'wordy' and thus more demanding. The biological problem-solving questions likely required advanced reading literacy as well as advanced understanding of complex biological concepts, and could not be answered by students' logic alone.

Higher quality questions were of higher importance for biology, and the questions of 'higher importance' were simultaneously targeted to evaluate higher cognitive levels of students as well as the importance of the students' knowledge for everyday life. The questions of higher importance used in our tool-testing-exam were complex and demanding – thus, hard to construct and likely shaped with less success (i.e., often affected by the students' logical reasoning during the problem solving tasks). Despite certain weaknesses, the questions designated as highly important for understanding biological processes and concepts (for students aged 13) represent quality questions within the present study, as they greatly encourage students' critical thinking.

Questions important for the curriculum (IC) demonstrated highly positive and significant correlation with the influence of questions on students' answers (Qin) (Table 3), suggesting that questions highly important for the curriculum may greatly be influenced by question shaping and intelligibility as well as by students' logical reasoning and learning paths.

The importance of questions for critical thinking (ICT) was positively correlated with the question quality (QQ) (Table 3), indicating that the questions of higher quality within the written exam encourage development of students' critical thinking.

Table 3. Spearman Rank Order Correlations

	SSR											
CL	-0.46	CL										
IP	-0.19	0.48	IP									
IL	-0.25	0.35	0.48	IL								
IC	0.43	-0.15	-0.09	0.18	IC							
ICT	0.05	-0.10	-0.10	0.18	0.40	ICT						
Qim	**-0.25**	**0.35**	**0.48**	**1.00**	**0.18**	**0.18**	**Qim**					
QS	0.52	-0.52	-0.26	-0.50	0.09	-0.06	**-0.50**	QS				
QI	0.35	-0.52	-0.46	-0.16	0.39	0.40	**-0.16**	0.39	QI			
SLR	0.05	-0.14	0.49	0.44	0.40	0.29	**0.44**	-0.06	0.17	SLR		
SLP	-0.16	0.43	0.28	0.18	0.09	-0.29	**0.18**	0.09	-0.21	-0.06	SLP	
Qin	**0.45**	**-0.15**	**0.11**	**0.18**	**0.70**	**0.17**	**0.18**	**0.39**	**0.39**	**0.40**	**0.39**	**Qin**
QQ	**0.01**	**-0.02**	**0.47**	**0.48**	**0.37**	**0.55**	**0.48**	**-0.10**	**0.13**	**0.73**	**-0.10**	**0.37**

SSR - student success rate of the exam
CL - cognitive level

MD pairwise deleted;
Bold correlations are significant
p < 0,05

Concordance as a measure of agreement between the evaluators'/teachers' opinions indicated a weaker concordance of the reasoning among assessors (average Fleiss kappa = 0.32) (Figure 2). There was a greater concordance among teachers regarding the assessment of the importance of questions (Kendall W = 0.53; ChiSq = 46.42; df = 22; p = 0.001) than regarding the assessment of the influence of question on students' answer (Kendall W = 0.31; ChiSq = 27.11; df = 22; p = 0.21). Significant concordance among the evaluators was recorded for the assessment of the question quality, influence of the students' logical reasoning on answering the question, and the importance of questions for critical thinking, curriculum and life (Figure 2). There was no significant concordance among the evaluators regarding the importance of questions for biology. It suggests that the evaluators disagree in their opinions, most likely because the key biological concepts and the respective conceptual framework are not clearly defined within the existing curriculum. Furthermore, there was no significant concordance among the evaluators regarding assessing the influence of question shape and intelligibility, and students' further learning path on answering the questions. It was again likely the result of lacking national standards and/or teachers' experience and/or a consequence of the low number of evaluators within this study.

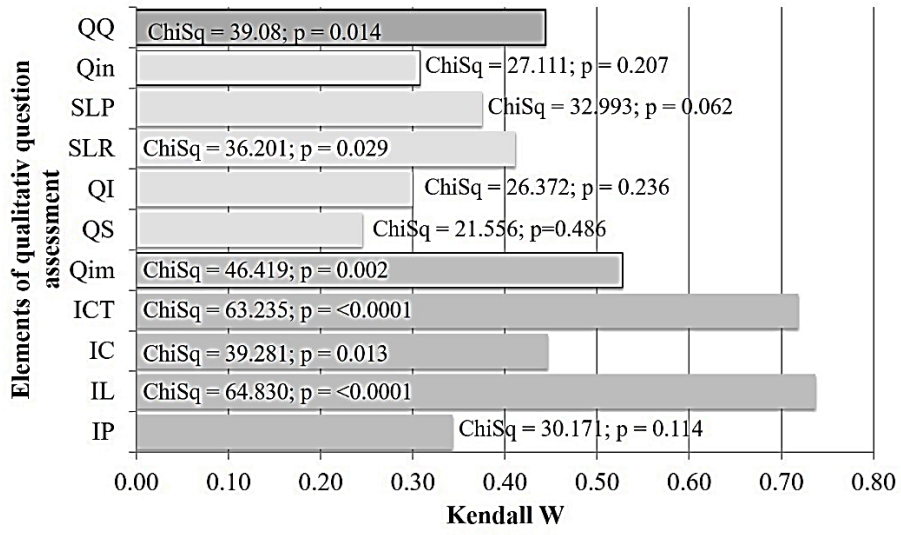


Figure 2 Concordance among teachers regarding the assessment

The results of qualitative question assessment may be helpful in order to get a better understanding of the percentage of answered questions and the student performance, respectively.

Out of 23 questions (Figure 3), 90% of students correctly answered on two acceptable questions - hard reproductive question 3 and easy conceptual question 6. Reproductive but difficult question number 19 had the worst success rate (the mean score of all students was less than the average score of the possible points). Poorly solved questions were questions 7 (15%) and 20 (14%), both moderately difficult and enquiring students' application of knowledge and conceptual understanding. Highly hard, medium quality question 21 was focused on checking the students' problem-solving ability, it had the highest score number, but was successfully answered by only 7% students.

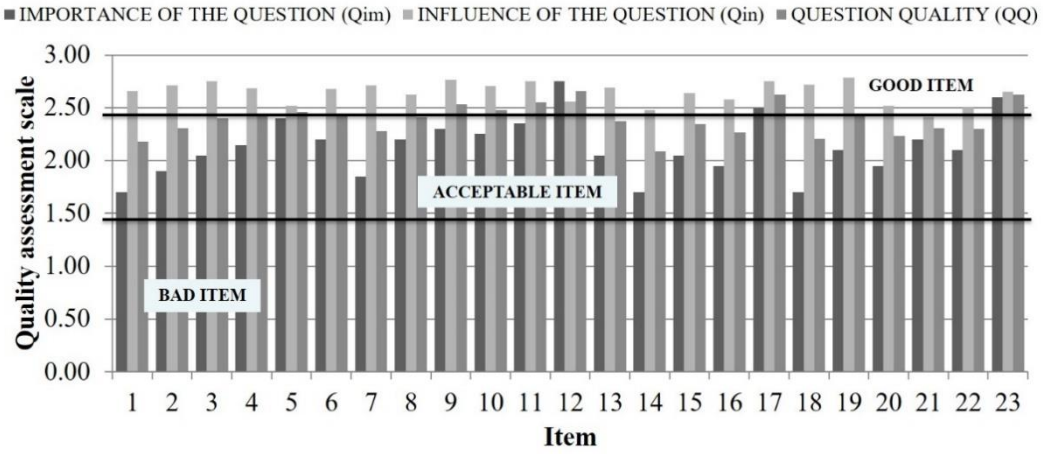


Figure 3. Comparison of points scored and index of item difficulty

Based on the assessments, there were no bad questions. Five questions could be designated as good (9, 11, 12, 17 i 23), and the rest as acceptable (Fig. 3). The quality of questions had likely low influence on students' answering, while only 3 questions (12, 17, 23), could be labelled as important.

According to the final question quality assessment done by averaging the scales' scores (Table 1), there were no statistically significant differences between the individual teachers' assessments (Kruskall-Wallis H = 0.25; df = 3; p = 0.97), and the teachers were relatively well-matched in their assessments (Kendall W = 0.44; ChiSq = 39.08; df = 22; p = 0.01). Authors of the questions were shown to be less self-critical in the self-evaluation than their peers, but this difference in the self-assessment was very low (8.6%) (Figure 4).

DISCUSSION AND CONCLUSIONS

Already during the initial application of our tool, it was noted that the critical assessment of the elements and criteria coincide with the results of psychometric question analysis (Radanović et al., 2010). Quality of the questions has lasting effects on teaching and learning, so the technical properties of the questions should be greatly considered by developers and practitioners (Dunbar et al. 2009). Discordance among the teachers' assessments confirms that the teachers are not prone to critically reflect on the questions they shape.

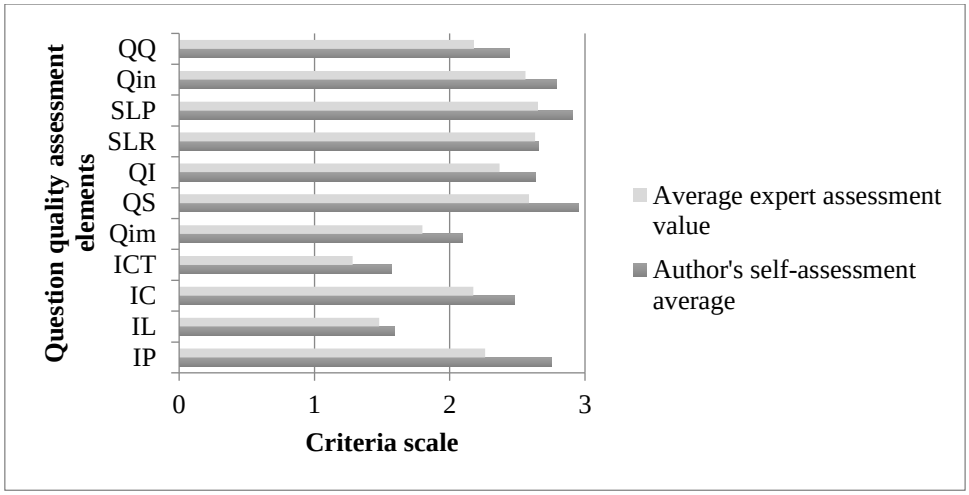


Figure 4. Comparison of author and teacher expert assessment

The teachers infrequently completed post-hoc statistical analyses of their tests (Marso & Pigge 1988) so a relatively simple quality analyses of their exams, based on the averaging valuation (i.e., consensus) among the selected elements and criteria for the question quality assessment, would provide plenty useful and relevant information on the overall question quality. More uniform teachers' assessments of the importance of questions might confirm the teachers' competence within the subject (i.e., biology), their knowledge and professional expertise. As the majority of open-ended items that are successfully tested for a higher cognitive level of knowledge, it is of utmost importance that the final say in deciding whether the item is effective in written evaluation must be given by the subject scientific basis (Begić et al., 2016), because according to Schmelzing et al. (2013) such issues have high content validity and potentially poorer inter-rater objectivity. The teachers' disagreement in the assessment of the influence of questions on students' answers could indicate an uneven teaching experience. Such result suggests that the teachers should necessarily continuously work on their own professional development (Gottheiner & Siegel 2012) to be able to focus well on setting the question quality standards (e.g., technical preparation of the questions, adaptation of the questions to the students, avoiding questions that demand high level of logical thinking, etc.). The teacher professional development should further help teachers to close the formative assessment cycle by addressing conceptions that are elicited with assessments (Gottheiner & Siegel 2012). Additionally, there is a need to develop the result analysis criteria for the exams, and a scientifically based approach to their assessment (Golovachyova et al. 2016). The tool we developed could be used for peer-evaluation as well as for self-assessment, but only if critically applied with the recommended delay of at least 2 weeks after the question preparation. Due to small number of teachers/evaluators (n = 4) in this study, our results indicate a certain trend, but to generalize our findings, our tool should be checked with a larger number of teachers and students' exams. The most important roles in the question quality assessment play the teachers' experience in the classroom as well as the overall experience in the question analysis. Therefore, it is very important to encourage teachers to collaborate in qualitative assessment of exam tasks.

ACKNOWLEDGEMENT

We thank Marijana Bastić and Ivanka Podrug for assessing the item quality.

REFERENCES

Begić, V., Bastić, M., & Radanović, I. (2016). Influence of students' biological knowledge in solving complex cognitive tasks. *Educ. biol.*, 2, 13-48.

Bellová, R., Melicherčíková, D., & Tomčík, P. (2017). Possible reasons for low scientific literacy of Slovak students in some natural science subjects. *Research in Science & Technological Education*, 1-17. https://doi.org/10.1080/02635143.2017.1367656 (accessed 16.12.2017)

Chang, A. (2014). *Statistics Toolkit (StatsToDo), Department of Obstetrics and Gynaecology*, the Chinese University of Hong Kong, available at https://www.statstodo.com/index.php, (accessed 24.01.2017).

Cohen, L., Manion, L., & Morrison, K. (2007). *Metode istrazivanja u obrazovanju* [Research methods in education]. Naklada Slap, Jastrebarsko.

Crooks, T. J. (1988) The Impact Of Classroom Evaluation Practices On Students, *Review of Educational Research, 58*(4): 438-481.

Downing, S.M. (2003). Validity: on the meaningful interpretation of assessment dana. *Medical Education, 37*(9), 830-837.

Dunbar, S. B., Koretz, D.M., & Hoover, H.D. (2009). Quality Control in the Development and Use of Performance Assessments. *Applied Measurement in Education, 4*, 289-303.

Golovachyova, V.N., Menlibekova, G.Zh., Abayeva, N.F., Ten, T.L., & Kogaya, G.D. (2016). Construction of Expert Knowledge Monitoring and Assessment System Based on Integral Method of Knowledge Evaluation. I*nternational Journal of Environmental and Science Education, 11*(9), 2539-2552.

Gottheiner, D. M., & Siegel, M. A. (2012). Experienced Middle School Science Teachers' Assessment Literacy: Investigating Knowledge of Students' Conceptions in Genetics and Ways to Shape Instruction. *Journal of Science Teacher Education, 23*(5), 531–557.

Hopkins, W.G. (2000). *A new view of statistics. Internet Society for Sport Science*, available at http://www.sportsci.org/resource/stats/ (accessed 15.11.2010).

MacCann, C., Roberts, R. D., Matthews, G., & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance-based Emotional Intelligence (EI) tests. *Personality and Individual Differences, 36*, 645–662.

Marso, R.N., & Pigge, F.L. (1988). *An Analysis of Teacher-Made Tests: Testing Practices, Cognitive Demands, and Item Construction Errors*. Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988). ED298174, 50.

OECD. 2015. PISA 2015 Results in Focus. http://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf. (accessed 13.9.2017).

Radanović, I., Ćurković, N., Bastić, M., Leniček, S., Furlan, Z., Španović, P., & Valjak-Porupski, M. (2010). *Qualitative analysis of Biology exams in primary school conducted in 2008.*, National centre for external evaluation of education, Zagreb, 111.

Radanović, I., Lukša, Ž., Garašić, D., Bastić, M., Marković, N., Furlan, Z., Dolenec, T., Begić, V., Kapov, S., Štiglić, N., & Petrač, T. (2011). *External evaluation exams in Biology in the eighth grade in school year 2010-2011. - The main trial*. National center for external evaluation of education, Zagreb, 114.

Radanović I., Lukša Ž., Pongrac Štimac Z., Garašić D., Bastić M., Kapov S., Kostanić LJ., Sertić Perić M., & Toljan M. (2017a). *Content and methodological analysis of state biology exam in the school year 2015./2016.* National centre for external evaluation of education, Zagreb, 212.

Radanović I., Lukša Ž., Begić V., Bastić M., Gotlibović G., Kapov S., Pavunec S., & Toljan M. (2017b). *Content and methodological analysis of state mature exams from Biology of School Years 2013./2014. i 2014./2015.* National centre for external evaluation of education, Zagreb, 101.